M. Krasnov A. Kiselev
G. Makarenko E. Shikin

Mathematical Analysis for Engineers

Volume 1

Mir Publishers Moscow

Курс высшей математики для инженеров

М. Краснов,
А. Киселев,
Г. Макаренко,
Е. Шикин

В двух томах

Том
1

Mathematical Analysis **for** Engineers

M. Krasnov
A. Kiselev
G. Makarenko
E. Shikin

In two volumes

Volume
1

Mir Publishers
Moscow

Translated from Russian
by Alexander Yastrebov

First published 1990

На английском языке

Printed in the Union of Soviet Socialist Republics

ISBN 5-03-000270-7
ISBN 5-03-000269-3

© Mir Publishers, 1989

Contents

Preface 9
Chapter 1 An Introduction to Analytic Geometry 11
1.1 Cartesian Coordinates 11
1.2 Elementary Problems of Analytic Geometry 14
1.3 Polar Coordinates 18
1.4 Second- and Third-Order Determinants 19
Chapter 2 Elements of Vector Algebra 24
2.1 Fixed Vectors and Free Vectors 24
2.2 Linear Operations on Vectors 26
2.3 Coordinates and Components of a Vector 30
2.4 Projection of a Vector onto an Axis 33
2.5 Scalar Product of Two Vectors 34
2.6 Vector Product of Two Vectors 39
2.7 Mixed Products of Three Vectors 43
Exercises 45
Answers 46
Chapter 3 The Line and the Plane 47
3.1 The Plane 47
3.2 Straight Line in a Plane 51
3.3 Straight Line in Three-Dimensional Space 55
Exercises 60
Answers 62
Chapter 4 Curves and Surfaces of the Second Order 63
4.1 Changing the Axes of Coordinates in a Plane 63
4.2 Curves of the Second Order 66
4.3 The Ellipse 67
4.4 The Hyperbola 71
4.5 The Parabola 77
4.6 Optical Properties of Curves of the Second Order 79
4.7 Classification of Curves of the Second Order 83
4.8 Surfaces of the Second Order 89
4.9 Classification of Surfaces 90
4.10 Standard Equations of Surfaces of the Second Order 95
Exercises 102
Answers 102

Chapter 5 Matrices. Determinants. Systems of Linear Equations 103
5.1 Matrices 103
5.2 Determinants 122
5.3 Inverse Matrices 133
5.4 Rank of a Matrix 139
5.5 Systems of Linear Equations 143
Exercises 165
Answers 167

Chapter 6 Linear Spaces and Linear Operators 168
6.1 The Concept of Linear Space 168
6.2 Linear Subspaces 170
6.3 Linearly Dependent Vectors 174
6.4 Basis and Dimension 175
6.5 Changing a Basis 181
6.6 Euclidean Spaces 183
6.7 Orthogonalization 185
6.8 Orthocompliments of Linear Subspaces 189
6.9 Unitary Spaces 191
6.10 Linear Mappings 192
6.11 Linear Operators 197
6.12 Matrices of Linear Operators 200
6.13 Eigenvalues and Eigenvectors 205
6.14 Adjoint Operators 209
6.15 Symmetric Operators 211
6.16 Quadratic Forms 213
6.17 Classification of Curves and Surfaces of the Second Order 221
Exercises 227
Answers 228

Chapter 7 An Introduction to Analysis 229
7.1 Basic Concepts 229
7.2 Sequences of Numbers 239
7.3 Functions of One Variable and Limits 247
7.4 Infinitesimals and Infinities 258
7.5 Operations on Limits 266
7.6 Continuous Functions. Continuity at a Point 272
7.7 Continuity on a Closed Interval 283
7.8 Comparison of Infinitesimals 288
7.9 Complex Numbers 294
Exercises 302
Answers 304

Chapter 8 Differential Calculus. Functions of One Variable 305
8.1 Derivatives and Differentials 305
8.2 Differentiation Rules 316
8.3 Differentiation of Composite and Inverse Functions 324
8.4 Derivatives and Differentials of Higher Orders 332
8.5 Mean Value Theorems 339
8.6 L'Hospital's Rule 344
8.7 Tests for Increase and Decrease of a Function on a Closed Interval and at a Point 349
8.8 Extrema of a Function. Maximum and Minimum of a Function on a Closed Interval 352
8.9 Investigating the Shape of a Curve. Points of Inflection 362
8.10 Asymptotes of a Curve 367
8.11 Curve Sketching 373
8.12 Approximate Solution of Equations 381
8.13 Taylor's Theorem 385
8.14 Vector Function of a Scalar Argument 396
Exercises 401
Answers 403
Chapter 9 Integral Calculus. The Indefinite Integral 409
9.1 Basic Concepts and Definitions 409
9.2 Methods of Integration 414
9.3 Integrating Rational Function 424
9.4 Integrals Involving Irrational Functions 435
9.5 Integrals Involving Trigonometric Functions 445
Exercises 450
Answers 453
Chapter 10 Integral Calculus. The Definite Integral 456
10.1 Basic Concepts and Definitions 456
10.2 Properties of the Definite Integral 461
10.3 Fundamental Theorems for Definite Integrals 467
10.4 Evaluating Definite Integrals 472
10.5 Computing Areas and Volumes by Integration 476
10.6 Computing Arc Lengths by Integration 488
10.7 Applications of the Definite Integral 495
10.8 Numerical Integration 498
Exercises 503
Answers 505
Chapter 11 Improper Integrals 506
11.1 Integrals with Infinite Limits of Integration 506
11.2 Integrals of Nonnegative Functions 511
11.3 Absolutely Convergent Improper Integrals 514

11.4 Cauchy Principal Value of the Improper Integrals 519
11.5 Improper Integrals of Unbounded Functions 520
11.6 Improper Integrals of Unbounded Nonnegative Functions. Convergence Tests 523
11.7 Cauchy Principal Value of the Improper Integral Involving Unbounded Functions 525
Exercises 526
Answers 527
Chapter 12 Functions of Several Variables 529
12.1 Basic Notions and Notation 529
12.2 Limits and Continuity 533
12.3 Partial Derivatives and Differentials 538
12.4 Derivatives of Composite Functions 545
12.5 Implicit Functions 550
12.6 Tangent Planes and Normal Lines to a Surface 555
12.7 Derivatives and Differentials of Higher Orders 558
12.8 Taylor's Theorem 562
12.9 Extrema of a Function of Several Variables 566
Exercises 580
Answers 583
Appendix I Elementary Functions 587
Index 596

Preface

This two-volume book was written for students of technical colleges who have had the usual mathematical training. It contains just enough information to continue with a wide variety of engineering disciplines. It covers analytic geometry and linear algebra, differential and integral calculus for functions of one and more variables, vector analysis, numerical and functional series (including Fourier series), ordinary differential equations, functions of a complex variable, Laplace and Fourier transforms, and equations of mathematical physics. This list itself demonstrates that the book covers the material for both a basic course in higher mathematics and several specialist sections that are important for applied problems. Hence, it may be used by a wide range of readers. Besides students in technical colleges and those starting a mathematics course, it may be found useful by engineers and scientists who wish to refresh their knowledge of some aspects of mathematics.

We tried to give the fundamental material concisely and without distracting detail. We concentrated on the presentation of the basic ideas of linear algebra and analysis to make it detailed and as comprehensible as possible. Mastery of these ideas is a requirement to understand the later material.

The many examples also serve this aim. The examples were written to help students with the mechanics of solving typical problems.

More than 600 diagrams are simple illustrations, clear enough to demonstrate the ideas and statements convincingly, and can be fairly easily reproduced.

We were conscious not to burden the course with scrupulous proofs for theorems which have little practical application. As a rule we chose the proof (marked in the text with special symbols) that was constructive in nature or explained fundamental ideas that had been introduced, showing how they work. This approach made it possible to devise algorithms for solving whole classes of important problems.

In addition to the examples, we have included several carefully selected problems and exercises (around 1000) which should be of interest to those pursuing an independent mathematics course. The problems have the form of moderately sized theorems. They are very simple but are good training for those learning the fundamental ideas.

Chapters 1-6, 26 and Appendix II were written by E. Shikin, Chapters 7-8, 11, 12, 17-21, 27, 28 and 29-32 by M. Krasnov, Chapters 9, 10, 13-16 by A. Kiselev, and Chapters 22-25 and Appendix I by G. Makarenko. There was no general editor, but each of the authors has read the chapters written by the colleagues, and so each chapter benefited from collective advice.

The Authors

Chapter 1

An Introduction to Analytic Geometry

1.1 Cartesian Coordinates

Coordinate axis. Let L be an arbitrary line. We may move along L in either of two directions. When the direction of moving is fixed, the line is said to be *directed.*

Definition. A directed line is called an *axis.*

The direction of an axis is indicated by an arrow (Fig. 1.1).

We fix on the axis L a point O and a line segment a of a unit length, called a unit distance (Fig. 1.2). Let M be a point on L. We associate with M a number x such that the value of x is equal to the positive distance between O and M if the direction of moving from O to M coincides with the direction of L, and to the negative distance otherwise (Fig. 1.3).

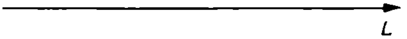


Fig. 1.1

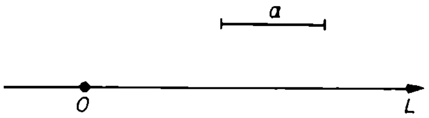


Fig. 1.2

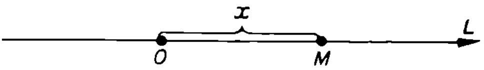


Fig. 1.3

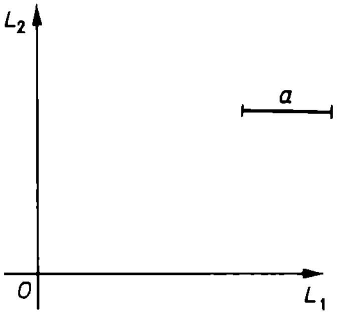


Fig. 1.4

Definition. The axis L with the *reference point* O and the *unit distance* a given on it is called the *coordinate axis;* the number x as defined above is said to be the *coordinate* of M.

In symbols, we write $M(x)$ to designate a point M whose coordinate is x.

Cartesian coordinates in a plane. Let O be a point in a plane. We draw through O two mutually perpendicular lines L_1 and L_2. Let us choose a direction for each line and a unit distance a which is the same for L_1 and L_2. Then L_1 and L_2 become coordinate axes with a common reference point O (Fig. 1.4).

We call one of the coordinate axes the x-axis or the *axis of abscissas* and the other one the y-axis or the *axis of ordinates* (Fig. 1.5). The point O is called the *origin of coordinates.*

Let M be a point in a plane as shown in Fig. 1.6. We drop from M two perpendiculars onto the coordinate axes, the points M_x and M_y being the projections of M on the x- and y-axes and associate with M an ordered pair (x, y) of numbers so that x is the coordinate of the point M_x and y is the coordinate of the point M_y.

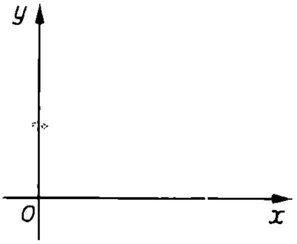


Fig. 1.5

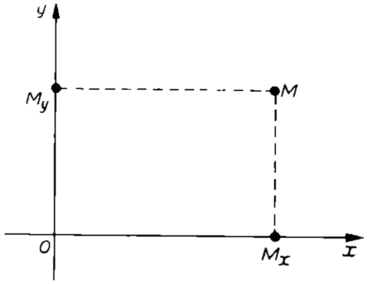
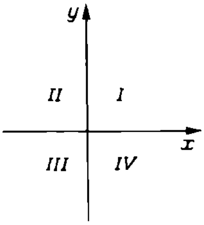


Fig. 1.6



	I	II	III	IV
x	+	−	−	+
y	+	+	−	−

Fig. 1.7

The numbers x and y are called the *Cartesian coordinates* of M, x being the *abscissa* and y being the *ordinate.*

In symbols, we write $M(x, y)$ to designate a point M whose coordinates are x and y.

For short, we shall refer to the frame of reference given above as a Cartesian coordinate system set up in a plane.

The coordinate axes divide a plane into four parts called *quadrants.* These are numbered as shown in Fig. 1.7 and the accompanying table.

Remark. The unit distances may be different for the two axes. Then the coordinate system is called *rectangular.*

Cartesian coordinates in three-dimensional space. Let O be a point in three-dimensional space. We draw through O three mutually perpendicular lines L_1, L_2 and L_3. We choose a direction for each line and a unit distance which is the same for L_1, L_2 and L_3. Then L_1, L_2 and L_3 become coordinate axes with a common reference point O (Fig. 1.8).

We call one of the axes the x-axis or the *axis of abscissas*, the second one the y-axis or the *axis of ordinates*, and the third one the z-axis or the *axis of applicates*. The point O is called the *origin of coordinates* (Fig. 1.9).

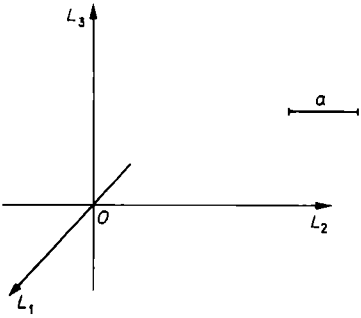


Fig. 1.8

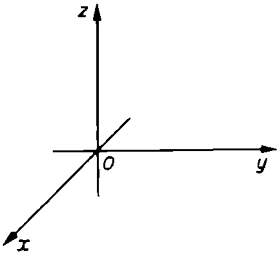


Fig. 1.9

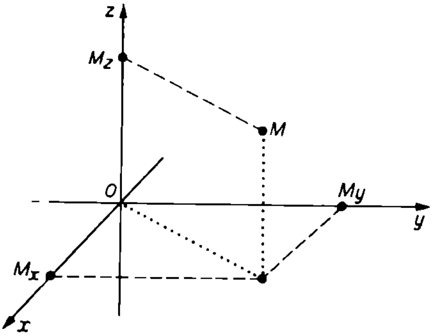


Fig. 1.10

Let M be an arbitrary point in three-dimensional space as shown in Fig. 1.10. We drop from M three perpendiculars onto the coordinate axes, the points M_x, M_y and M_z being the projections of M on the x-, y- and z-axes and associate with M an ordered triple (x, y, z) of numbers, so that x is the coordinate of the point M_x, y is the coordinate of the point M_y and z is the coordinate of the point M_z.

The numbers x, y and z are said to be the *Cartesian coordinates* of M; x, y and z are called the *abscissa, ordinate* and *applicate* of the point M, respectively.

In symbols, we write $M(x, y, z)$ to designate a point M whose coordinates are x, y and z.

Thus we have set up a Cartesian coordinate system in three-dimensional space.

Definition. A plane determined by any pair of coordinate axes is called a *coordinate plane.*

There are three coordinate planes in three-dimensional space, namely the xy-plane, the xz-plane and the yz-plane. These planes divide the space into eight parts called *octants.*

1.2 Elementary Problems of Analytic Geometry

Distance formulas. Let $M_1(x_1)$ and $M_2(x_2)$ be points on a coordinate axis. Then the distance d between $M_1(x_1)$ and $M_2(x_2)$ is given by

$$d = d(M_1, M_2) = |x_2 - x_1|.$$

Let there be given a Cartesian coordinate system in an xy-plane. Then the distance between any two points $M_1(x_1, y_1)$ and $M_2(x_2, y_2)$ is given by

$$d = d(M_1, M_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

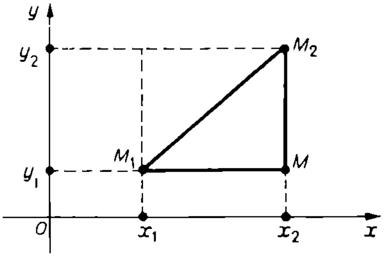


Fig. 1.11

◄ Consider the right triangle MM_1M_2 (Fig. 1.11). The theorem of Pythagoras gives $|M_1M_2|^2 = |M_1M|^2 + |MM_2|^2$. Since the distance between M_1 and M_2 equals the length of the segment M_1M_2 and $|M_1M| = |x_2 - x_1|$, $|MM_2| = |y_2 - y_1|$, we have

$$d^2 = |x_2 - x_1|^2 + |y_2 - y_1|^2.$$

Notice that $|x_2 - x_1|^2 = (x_2 - x_1)^2$ and $|y_2 - y_1|^2 = (y_2 - y_1)^2$. Then extracting the square root from d^2, we get the desired formula. ►

Remark. In three-dimensional space the distance between $M_1(x_1, y_1, z_1)$ and $M_2(x_2, y_2, z_2)$ is

$$d = d(M_1, M_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}\,.$$

(Show this.)

Examples. (1) Write the equation of a circle with radius r and centre at the point $P(a, b)$.

◄ Let $M(x, y)$ be a point of a circle (Fig. 1.12). Then $|MP| = r$. Since $|MP|$ is the distance between M and P we have

$$|MP| = r = \sqrt{(x - a)^2 + (y - b)^2}\,.$$

Squaring this equation, we get

$$(x - a)^2 + (y - b)^2 = r^2.$$

This is the desired *equation of a circle.* ►

(2) Let $F_l(-c, 0)$ and $F_r(c, 0)$ be points in a plane and a $(a > c \geqslant 0)$ be a given number. Find the condition to be satisfied by coordinates (x, y) of a point M for the sum of the distances between M and F_l and between M and F_r to be equal to $2a$.

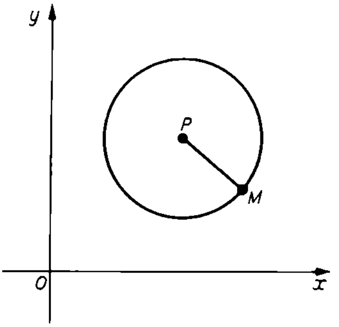


Fig. 1.12

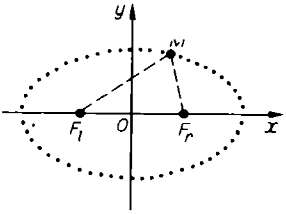


Fig. 1.13

◄ Let us find the distances between M and F_l and between M and F_r (Fig. 1.13). We have

$$|MF_l| = \sqrt{(x + c)^2 + y^2} \quad \text{and} \quad |MF_r| = \sqrt{(x - c)^2 + y^2}\,.$$

Whence

$$\sqrt{(x + c)^2 + y^2} + \sqrt{(x - c)^2 + y^2} = 2a.$$

Transpose the second radical to the right:

$$\sqrt{(x + c)^2 + y^2} = 2a - \sqrt{(x - c)^2 + y^2}\,.$$

Then, squaring and simplifying both sides of the equation, we get

$$a\sqrt{(x-c)^2+y^2} = a^2 - cx.$$

Squaring and further simplifying both sides of the above equation, we obtain

$$(a^2-c^2)x^2 + a^2y^2 = a^2(a^2-c^2).$$

Setting $b^2 = a^2 - c^2$ and dividing both sides by a^2b^2, we arrive at the *equation of an ellipse* (see Chap. 4)

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

This is the condition we have sought for. ►

Division of a line segment in a given ratio. Let $M_1(x_1, y_1)$ and $M_2(x_2, y_2)$ be two distinct points in a plane. Let a point $M(x, y)$ lie on the line segment M_1M_2 and divide M_1M_2 in the ratio $\lambda_1:\lambda_2$, i.e.,

$$\frac{|M_1M|}{|MM_2|} = \frac{\lambda_1}{\lambda_2}.$$

Represent the coordinates (x, y) of M in terms of the coordinates of M_1 and M_2 and the numbers λ_1 and λ_2.

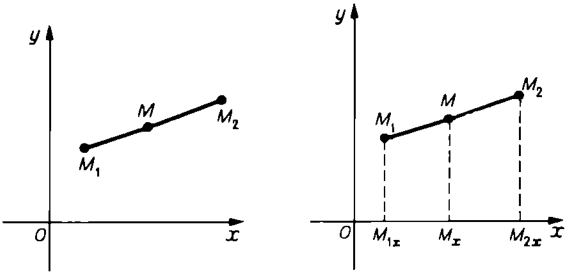


Fig. 1.14

◄ Suppose that the segment is not parallel to the y-axis (Fig. 1.14). Then

$$\frac{|M_1M|}{|MM_2|} = \frac{|M_{1x}M_x|}{|M_xM_{2x}|}.$$

Since $|M_{1x}M_x| = |x_1 - x|$ and $|M_xM_{2x}| = |x - x_2|$ we have

$$\frac{|x_1 - x|}{|x - x_2|} = \frac{\lambda_1}{\lambda_2}.$$

The point M lies between M_1 and M_2. Hence, there holds either $x_1 < x < x_2$ or $x_1 > x > x_2$. This implies that the differences $x_1 - x$ and $x - x_2$ are always of the same sign. Thus we may write

$$\frac{x_1 - x}{x - x_2} = \frac{\lambda_1}{\lambda_2}.$$

Hence

$$x = \frac{\lambda_2 x_1 + \lambda_1 x_2}{\lambda_1 + \lambda_2}. \qquad (*)$$

When the segment M_1M_2 is parallel to the y-axis, we have $x_1 = x_2 = x$. Notice that this result immediately follows from (*) if we set $x_1 = x_2$.

Similar reasoning yields

$$y = \frac{\lambda_2 y_1 + \lambda_1 y_2}{\lambda_1 + \lambda_2}. \quad ►$$

Example. Find the coordinates of the centre of gravity M of the triangle with vertices at $M_1(x_1, y_1)$, $M_2(x_2, y_2)$ and $M_3(x_3, y_3)$ (Fig. 1.15).

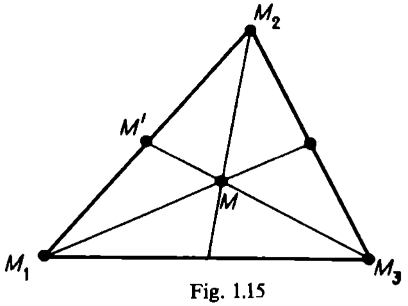


Fig. 1.15

◄ Recall that in any triangle the centre of gravity and the point of intersection of medians coincide so that M divides each median in the ratio 2:1 reckoning from the corresponding vertex. Thus, the coordinates of M are

$$x = \frac{1x_3 + 2x'}{2 + 1} \quad \text{and} \quad y = \frac{1y_3 + 2y'}{2 + 1},$$

where x' and y' are the coordinates of the point M' of the median M_3M'.

Since M' is the mid-point of M_1M_2, we have

$$x' = \frac{1x_1 + 1x_2}{1 + 1} \quad \text{and} \quad y' = \frac{1y_1 + 1y_2}{1 + 1}.$$

Substituting these relations into the formulas for x and y, we arrive at the desired result

$$x = \frac{x_1 + x_2 + x_3}{3} \quad \text{and} \quad y = \frac{y_1 + y_2 + y_3}{3}. \quad ►$$

Remark. Let $M(x, y, z)$ divide a segment joining $M_1(x_1, y_1, z_1)$ and $M_2(x_2, y_2, z_2)$ in the ratio $\lambda_1 : \lambda_2$. Then

$$x = \frac{\lambda_2 x_1 + \lambda_1 x_2}{\lambda_1 + \lambda_2}, \quad y = \frac{\lambda_2 y_1 + \lambda_1 y_2}{\lambda_1 + \lambda_2}, \quad z = \frac{\lambda_2 z_1 + \lambda_1 z_2}{\lambda_1 + \lambda_2}.$$

1.3 Polar Coordinates

Consider an axis L in a plane and a point O on L (Fig. 1.16). Let M be a point distinct from O as shown in Fig. 1.17. The number r is the distance between O and M and φ is the angle between the positive direction of L and OM measured counterclockwise. We easily see that the position of M in the plane is uniquely defined by the values of r and φ.

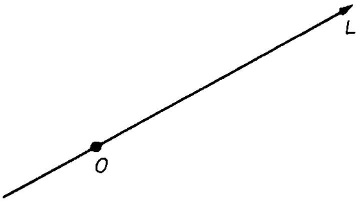


Fig. 1.16

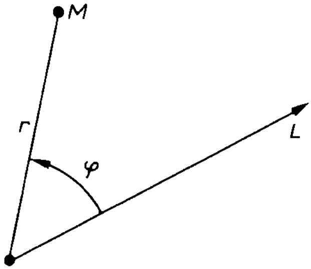


Fig. 1.17

The ordered pair (r, φ) of numbers represents the *polar coordinates* of M. The numbers r and φ are called the *polar radius* and the *polar angle,* respectively.

The point O is referred to as the *pole* and the axis L as the *polar axis.*

It is clear that $r > 0$ and $0 \leqslant \varphi < 2\pi$.

When M coincides with the pole, $r = 0$. In this case the polar angle is not defined.

Therefore we may set another coordinate system in a plane, namely the *polar coordinate system.*

The Cartesian coordinate system is said to be *compatible* with a given polar system if the origin O is the pole, the x-axis is the polar axis and the y-axis makes an angle of $+\frac{\pi}{2}$ with the x-axis (Fig. 1.18). Then the relations between the Cartesian coordinates and the polar coordinates are given by the formulas:

$$\cos\varphi = \frac{x}{r} \quad \text{and} \quad \sin\varphi = \frac{y}{r},$$

or

$$x = r\cos\varphi \quad \text{and} \quad y = r\sin\varphi,$$

where $r = \sqrt{x^2 + y^2}$.

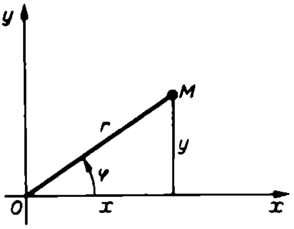


Fig. 1.18

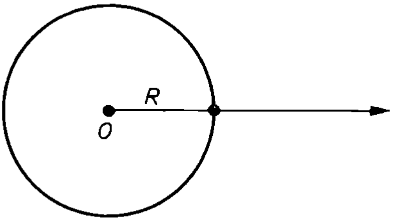


Fig. 1.19

Example. Let R ($R > 0$) be a given number. The set of points whose polar coordinates (r, φ) satisfy the condition

$$r = R$$

is a circle with radius R and centre at the pole (Fig. 1.19).

1.4 Second- and Third-Order Determinants

We are given four numbers a_{11}, a_{12}, a_{21}, a_{22} (these are read *a* one-one, *a* one-two, *a* two-one, *a* two-two).

The *second-order determinant* is the number

$$a_{11}a_{22} - a_{12}a_{21}.$$

In symbols, we write

$$\Delta = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21} = \det |a_{ij}| \quad (i = 1,\ 2;\ j = 1,\ 2) \tag{1.1}$$

to designate the second-order determinant.

The numbers a_{11}, a_{12}, a_{21} and a_{22} are called *elements* of Δ. The pairs a_{11}, a_{12} and a_{21}, a_{22} of elements are referred to as the *first* and the *second rows* of Δ, respectively; a_{11}, a_{21} and a_{12}, a_{22} as the *first* and the *second columns* of Δ, respectively; a_{11}, a_{22} as the *principal* (or *positive*) *diagonal* of Δ; and a_{12}, a_{21} as the *secondary* (or *negative*) *diagonal* of Δ (see Fig. 1.20).

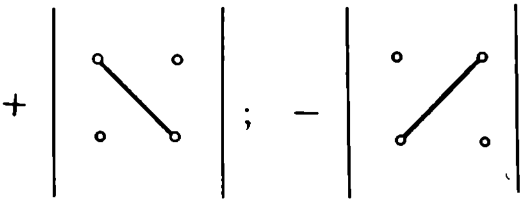


Fig. 1.20

Therefore the value of the second-order determinant is equal to the product of elements on the principal diagonal minus the product of elements on the secondary diagonal.

Example. Compute the value of

$$\begin{vmatrix} -1 & 3 \\ -2 & 4 \end{vmatrix}.$$

◄ Using (1.1), we get

$$\begin{vmatrix} -1 & 3 \\ -2 & 4 \end{vmatrix} = -4 + 6 = 2. \quad \blacktriangleright$$

We encounter the second-order determinants when dealing with systems of linear equations in two unknowns

$$\begin{cases} a_{11}x + a_{12}y = b_1 \\ a_{21}x + a_{22}y = b_2 \end{cases} \left| \begin{array}{l} a_{22} \\ -a_{12} \end{array} \right. .$$

Provided that $\Delta = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0$, eliminating the unknowns gives

$$x = \frac{b_1a_{22} - b_2a_{12}}{a_{11}a_{22} - a_{12}a_{21}} \quad \text{and} \quad y = \frac{b_2a_{11} - b_1a_{21}}{a_{11}a_{22} - a_{12}a_{21}}$$

or, as quotients of determinants,

$$x = \frac{\begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}} \quad \text{and} \quad y = \frac{\begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}}.$$

Let there be given nine numbers a_{ij} $(i = 1, 2, 3;\ j = 1, 2, 3)$.

The *third-order determinant*

$$\Delta = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \det |a_{ij}| \qquad .$$

is the number whose value is computed as

$$\begin{aligned} \Delta = {} & a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{21}a_{32}a_{13} \\ & - a_{13}a_{22}a_{31} - a_{21}a_{12}a_{33} - a_{11}a_{23}a_{32}. \end{aligned} \tag{1.2}$$

The first subscript i of a_{ij} refers to the row, the second subscript j refers to the column.

The triples of elements a_{11}, a_{22}, a_{33} and a_{13}, a_{22}, a_{31} are called the principal (or positive) diagonal and the secondary (or negative) diagonal of Δ, respectively.

To tackle a question of signs in (1.2) we mention that the three positive terms a_{11}, a_{22}, a_{33}; a_{12}, a_{23}, a_{31} and a_{21}, a_{32}, a_{13} correspond to the principal diagonal and to the vertices of the two triangles whose bases are parallel to the principal diagonal. Similarly, the secondary diagonal and the vertices of the two triangles whose bases are parallel to the secondary diagonal correspond to the three negative terms in (1.2), as shown in Fig. 1.21 and Fig. 1.22.

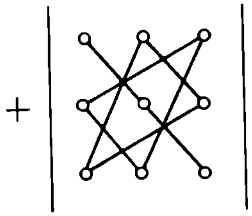


Fig. 1.21

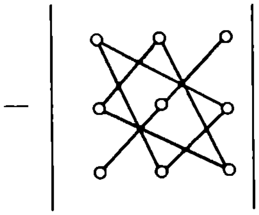


Fig. 1.22

This consideration lays down a convenient procedure for computing the value of the third-order determinant, called the *rule of a triangle.*

Example. Compute the value of

$$\Delta = \begin{vmatrix} 1 & 0 & -1 \\ 2 & 4 & 3 \\ 3 & -1 & 6 \end{vmatrix}.$$

◄ The rule of a triangle gives

$$\Delta = 1 \times 4 \times 6 + (-1) \times 2 \times (-1) + 0 \times 3 \times 3 - (-1) \times 3 \times 4$$
$$-0 \times 2 \times 6 - 1 \times (-1) \times 3 = 24 + 2 + 0 + 12 - 0 + 3 = 41. ►$$

Now we shall turn our attention to some properties of the second- and the third-order determinants easily verified by applying formulas (1.1) and (1.2).

Properties of determinants. (1) The value of the determinant is unchanged if the rows and the columns are transposed, i.e.,

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = \begin{vmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{vmatrix},$$

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{vmatrix}.$$

(2) The determinant reverses its sign if any two rows (or any two columns) are interchanged.

(3) The factor common to all elements of some row (or some column) may be taken outside the determinant, i.e.,

$$\begin{vmatrix} ka_{11} & ka_{12} & ka_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} a_{11} & ka_{12} & a_{13} \\ a_{21} & ka_{22} & a_{23} \\ a_{31} & ka_{32} & a_{33} \end{vmatrix} = k \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}.$$

The following three properties are consequences of Properties 1-3. We can also verify these properties by directly applying formulas (1.1) and (1.2).

(4) The determinant is zero if it contains two identical rows (or two identical columns).

(5) The determinant vanishes if all elements of some row (or some column) are zeros.

(6) The determinant is zero if it contains two proportional rows (or two proportional columns).

We shall outline another approach to computation of the third-order determinant

$$\Delta = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}.$$

The *minor* M_{ij} of Δ is the second-order determinant obtained by deleting from Δ the row and the column which intersect in a_{ij}.

For example, the minor M_{23} of Δ is the second-order determinant

$$M_{23} = \begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix}.$$

The *cofactor* A_{ij} of the element a_{ij} in Δ is the minor M_{ij} if the sum $(i + j)$ of the numbers of the ith row and the jth column is even, and the minor M_{ij} multiplied by (-1) otherwise, i.e.,

$$A_{ij} = (-1)^{i+j} M_{ij},$$

where $(-1)^{i+j}$ is the position sign.

Theorem 1.1. *The third-order determinant Δ is equal to the sum of products of elements in some row (or some column) of Δ multiplied by their respective cofactors so that the following expansions*

$$\Delta = a_{i1}A_{i1} + a_{i2}A_{i2} + a_{i3}A_{i3} \quad (i = 1, 2, 3), \tag{1.3}$$

$$\Delta = a_{1j}A_{1j} + a_{2j}A_{2j} + a_{3j}A_{3j} \quad (j = 1, 2, 3) \tag{1.4}$$

hold.

By way of illustration we show that

$$\Delta = a_{11}A_{11} + a_{12}A_{12} + a_{13}A_{13}.$$

◄ Applying formula (1.3), we obtain

$$\Delta = a_{11}(a_{22}a_{33} - a_{32}a_{23}) + a_{12}(a_{23}a_{31} - a_{21}a_{33}) + a_{13}(a_{21}a_{32} - a_{31}a_{22})$$

$$= a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

$$= a_{11}M_{11} - a_{12}M_{12} + a_{13}M_{13} = a_{11}A_{11} + a_{12}A_{12} + a_{13}A_{13}.$$ ►

Formulas (1.3) and (1.4) represent the *expansion of the determinant with respect to the ith row* and the *expansion of the determinant with respect to the jth column,* respectively.

Example. Compute the value of

$$\Delta = \begin{vmatrix} 1 & 0 & -1 \\ 2 & 4 & 3 \\ 3 & -1 & 6 \end{vmatrix}.$$

◄ Expanding the determinant with respect to the first row, we get

$$\Delta = 1\begin{vmatrix} 4 & 3 \\ -1 & 6 \end{vmatrix} - 0\begin{vmatrix} 2 & 3 \\ 3 & 6 \end{vmatrix} - 1\begin{vmatrix} 2 & 4 \\ 3 & -1 \end{vmatrix}$$

$$= (24 + 3) + 0 + (2 + 12) = 41.$$ ►

Chapter 2

Elements of Vector Algebra

2.1 Fixed Vectors and Free Vectors

Let A and B be two distinct points. We may move along a line segment joining A and B in either of two directions. Suppose we start moving at A and end at B. This gives rise to a *directed segment AB.* When moving from B to A, we get a *directed segment BA.*

Directed segments are frequently called *fixed vectors.*

The direction of a fixed vector (or a directed segment) is indicated pictorially by an arrow (Fig. 2.1).

We denote a fixed vector by a symbol, say $\overrightarrow{AB}$, with the understanding that the first letter in a pair refers to the initial point of the fixed vector and the second letter to the terminal point.

A fixed vector is said to be *zero* if its initial and terminal points coincide.

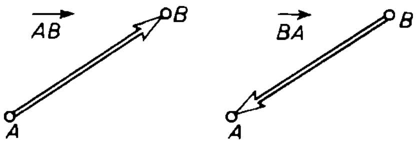


Fig. 2.1

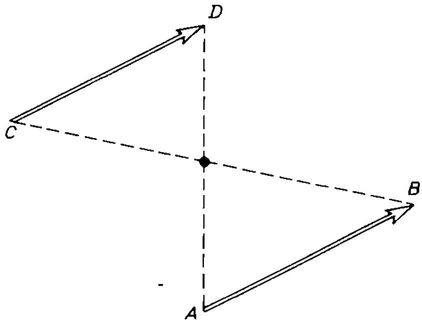


Fig. 2.2

Definition. The fixed vectors $\overrightarrow{AB}$ and $\overrightarrow{CD}$ are said to be *equivalent* if the mid-points of the segments AD and BC coincide (Fig. 2.2).

In symbols, we write $\overrightarrow{AB} = \overrightarrow{CD}$ to signify the fact that $\overrightarrow{AB}$ and $\overrightarrow{CD}$ are two equivalent fixed vectors.

Notice that when $\overrightarrow{AB}$ and $\overrightarrow{CD}$ do not lie in the same line, we may introduce a definition equivalent to the given above, namely $\overrightarrow{AB}$ and $\overrightarrow{CD}$ are equivalent if the quadrilateral *ABCD* is a parallelogram. Hence all equivalent fixed vectors are of the same length.

By way of illustration we consider the square and the fixed vectors shown in Fig. 2.3. It is clear that $\overrightarrow{AB}$ and $\overrightarrow{DC}$ are equivalent while $\overrightarrow{BC}$ and $\overrightarrow{DA}$ are not.

Equivalent fixed vectors obey the following laws.

(1) Any fixed vector is equivalent to itself, i.e., $\overrightarrow{AB} = \overrightarrow{AB}$.

(2) If $\overrightarrow{AB} = \overrightarrow{CD}$ then $\overrightarrow{CD} = \overrightarrow{AB}$.

(3) If $\overrightarrow{AB} = \overrightarrow{CD}$ and $\overrightarrow{CD} = \overrightarrow{EF}$ then $\overrightarrow{AB} = \overrightarrow{EF}$.

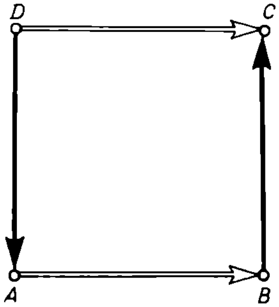


Fig. 2.3

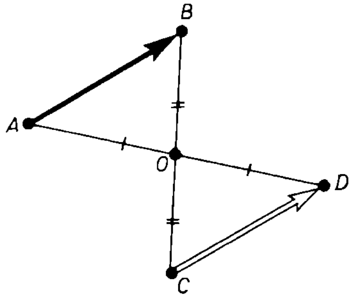


Fig. 2.4

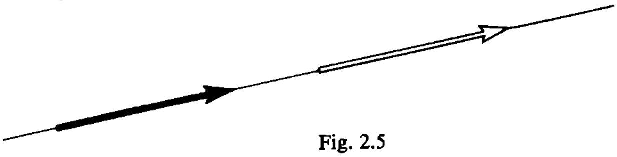


Fig. 2.5

Let $\overrightarrow{AB}$ be a given fixed vector and *C* be an arbitrary point. By virtue of the above definition we can always find a point *D* (Fig. 2.4) such that

$$\overrightarrow{CD} = \overrightarrow{AB}.$$

Hence, to any point we may apply a fixed vector equivalent to the given one.

Now we shall turn our attention to *free vectors*, i.e, to vectors which may be made to start at an arbitrary point. In other words, a free vector is a vector which may be moved rigidly parallel to itself.

We clearly see that a given fixed vector $\overrightarrow{AB}$ uniquely defines a free vector which starts at *A* and ends at *B*.

Free vectors whose initial points lie on the line determined by a given (nonzero) fixed vector are thought of as *sliding vectors* (Fig. 2.5).

Fixed vectors and sliding vectors are widely used in theoretical mechanics.

We shall denote free vectors by bold face small letters **a**, **b**, **c**, ... and the zero vector by **0**.

Let **a** be a free vector and A be a point. Then there exists only one point B such that

$$\overrightarrow{AB} = \mathbf{a}.$$

Hence, the operation of moving a given free vector to a specific point gives rise to a unique fixed vector $\overrightarrow{AB}$. This operation is referred to as applying a free vector **a** to a point A.

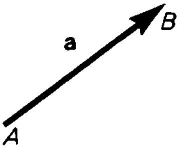


Fig. 2.6

Notice that applying a given free vector to different points, we obtain fixed vectors which are equivalent and, consequently, of the same length. Therefore we may speak of the *length of a free vector*, setting it equal to the length of the corresponding fixed vector. The length of the zero vector is equal to zero. We shall denote the length of a free vector **a** by $|\mathbf{a}|$.

Notice that if $\mathbf{a} = \mathbf{b}$ then $|\mathbf{a}| = |\mathbf{b}|$ and the converse is not true.

2.2 Linear Operations on Vectors

Addition of vectors. Let there be given two vectors **a** and **b**. Apply the vector **a** to some point O so that $\overrightarrow{OA} = \mathbf{a}$ and then apply the vector **b** to the point A, as shown in Fig. 2.7. The resultant vector $\overrightarrow{OB}$ is the *sum* of **a** and **b** denoted by $\mathbf{a} + \mathbf{b}$. This rule of addition of two vectors is known as the *triangle law*.

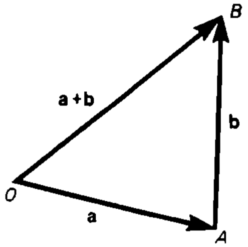


Fig. 2.7

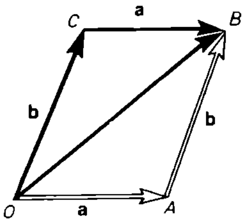


Fig. 2.8

It is easy to see that addition of two vectors obeys the *commutative law,* i.e., there holds

$$\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}.$$

for any two vectors $\mathbf{a}$ and $\mathbf{b}$ (Fig. 2.8).

Apply $\mathbf{a}$ and $\mathbf{b}$ to some point O and construct a parallelogram with sides $\mathbf{a}$ and $\mathbf{b}$, as shown in Fig. 2.9. Then the vector $\overrightarrow{OB}$ which starts at O and ends at the vertex opposite to O represents the sum $\mathbf{a} + \mathbf{b}$ (or $\mathbf{b} + \mathbf{a}$). This rule of addition is called the *parallelogram law.*

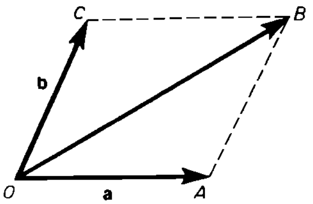


Fig. 2.9

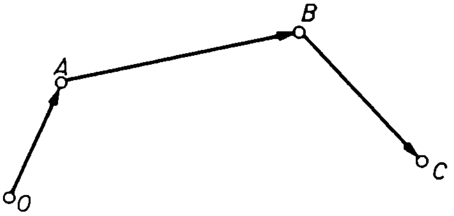


Fig. 2.10

Now we consider three vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ and proceed as follows. First, we apply $\mathbf{a}$ to some point O so that $\overrightarrow{OA} = \mathbf{a}$. Then we apply $\mathbf{b}$ to A so that $\overrightarrow{AB} = \mathbf{b}$. And finally, we apply $\mathbf{c}$ to B so that $\overrightarrow{BC} = \mathbf{c}$. Thus we obtain a polygonal line of three vectors as shown in Fig. 2.10.

Using the triangle law, we obtain $\overrightarrow{OB} = \mathbf{a} + \mathbf{b}$ and $\overrightarrow{OC} = (\mathbf{a} + \mathbf{b}) + \mathbf{c}$ (Fig. 2.11).

On the other hand, we have $\overrightarrow{AC} = \mathbf{b} + \mathbf{c}$ and $\overrightarrow{OC} = \mathbf{a} + (\mathbf{b} + \mathbf{c})$ (Fig. 2.12). Hence, we conclude that for any three vectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ there holds

$$(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c}).$$

Thus addition of three vectors obeys the *associative* law and we may write the sum of $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ as

$$\mathbf{a} + \mathbf{c} + \mathbf{b}.$$

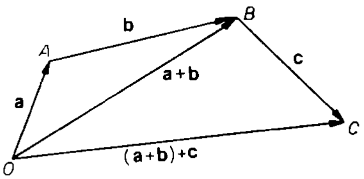


Fig. 2.11

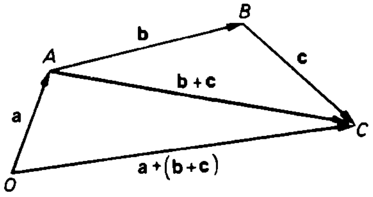


Fig. 2.12

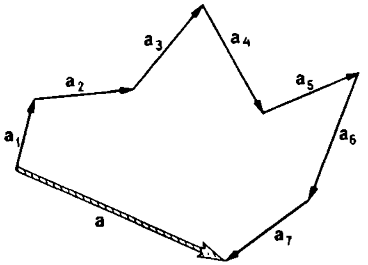


Fig. 2.13

Similarly, we define the sum of any number of vectors as a vector which starts at the initial point of the first summand and ends at the terminal point of the last summand as depicted in Fig. 2.13 for seven vectors. The sum of seven vectors is

$$\mathbf{a} = \mathbf{a}_1 + \mathbf{a}_2 + \mathbf{a}_3 + \mathbf{a}_4 + \mathbf{a}_5 + \mathbf{a}_6 + \mathbf{a}_7.$$

This rule of addition may be referred to as the *law of closure of a polygonal line to complete a polygon.*

Example. Find the sum of six vectors each of which starts at the centre and ends at a vertex of a regular hexagon (Fig. 2.14).

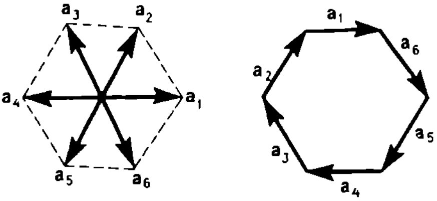


Fig. 2.14

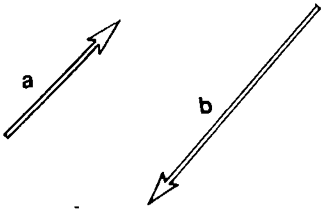


Fig. 2.15

◄ The law of closure of a polygonal line yields

$$\mathbf{a}_1 + \mathbf{a}_2 + \mathbf{a}_3 + \mathbf{a}_4 + \mathbf{a}_5 + \mathbf{a}_6 = \mathbf{0}. \blacktriangleright$$

Multiplication of a vector by a scalar. Definition. Free vectors **a** and **b** are said to be *collinear* if their corresponding fixed vectors lie on parallel lines (Fig. 2.15) or on the same line.

In symbols, we write $\mathbf{a} \parallel \mathbf{b}$ to denote collinear vectors.

Remark. It follows from the definition that **a** and **b** are always collinear when either **a** or **b** is a zero vector.

Let **a** and **b** be two collinear vectors applied to a point O so that $\overrightarrow{OA} = \mathbf{a}$ and $\overrightarrow{OB} = \mathbf{b}$. Then the points O, A and B lie on the same line. The point O may lie either outside the segment joining A and B or between A and B as shown in Fig. 2.16. The vectors **a** and **b** are of the same direction in the first case and of the opposite directions in the second one.

The vectors **a** and **b** are said to be *equivalent* if they are of the *same length* and of the *same direction.*

Let **a** be a vector and λ be a scalar.

Definition. The product of a vector **a** by a scalar λ is the vector **b** such that

(a) $|\mathbf{b}| = |\lambda| \cdot |\mathbf{a}|$;

(b) **a** and **b** are of the same direction if $\lambda > 0$ and of opposite directions when $\lambda < 0$.

In symbols, we write $\mathbf{b} = \lambda\mathbf{a}$.

When $\lambda = 0$ we get $\lambda\mathbf{a} = \mathbf{0}$.

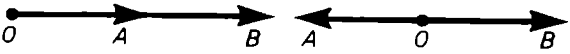


Fig. 2.16

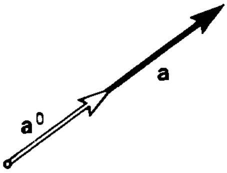


Fig. 2.17

Therefore the vectors **a** and $\mathbf{b} = \lambda\mathbf{a}$ are collinear by definition. The converse is also true, namely if **a** $(\mathbf{a} \neq \mathbf{0})$ and **b** are collinear vectors, then there exists a scalar λ such that $\mathbf{b} = \lambda\mathbf{a}$.

Basic properties of multiplication of a vector by a scalar are

(1) $(\lambda + \mu)\mathbf{a} = \lambda\mathbf{a} + \mu\mathbf{a}$.

(2) $\lambda(\mu\mathbf{a}) = (\lambda\mu)\mathbf{a}$.

(3) $\lambda(\mathbf{a} + \mathbf{b}) = \lambda\mathbf{a} + \lambda\mathbf{b}$.

(Here λ and μ are any real numbers, and **a** and **b** are any vectors.)

Definition. A vector $\mathbf{a}^0$ of a unit length $(|\mathbf{a}^0| = 1)$ is called a *unit vector.*

If $\mathbf{a} \neq \mathbf{0}$ then the vector

$$\mathbf{a}^0 = \frac{1}{|\mathbf{a}|}\mathbf{a} = \frac{\mathbf{a}}{|\mathbf{a}|}$$

is a unit vector of the same direction as **a** (Fig. 2.17).

2.3 Coordinates and Components of a Vector

Let a Cartesian coordinate system with the origin O be given in three-dimensional space.

Given a point M in the space, we may find three points P, Q and R called projections of M onto the coordinate axes (Fig. 2.18). Conversely, given three points P, Q and R on the coordinate axes, there exists a unique point M such that P, Q and R are projections of M onto the coordinate axes. Hence, the position of M in the space is uniquely defined by projections P, Q and R of M onto the coordinate axes and vice versa.

Recall that positions of P, Q and R on the coordinate axes are identified by their respective coordinates x, y and z. There is a one-to-one correspondence between M and the ordered triple (x, y, z) of numbers called the Cartesian coordinates of M.

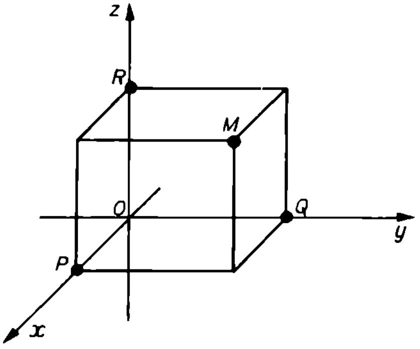


Fig. 2.18

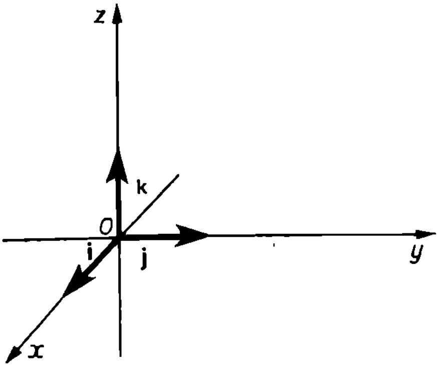


Fig. 2.19

Let **i**, **j**, **k** be three unit vectors located at the origin O of a given Cartesian coordinate system and pointing in the positive directions of the x-, y- and z-axes, respectively, as shown in Fig. 2.19. We consider a vector **a** which starts at the origin O of coordinates and ends at the point A. Let us draw through A planes perpendicular to coordinate axes. These planes intersect the x-, y- and z-axes at the points P, Q and R, respectively. It follows from Fig. 2.20 that

$$\mathbf{a} = \overrightarrow{OP} + \overrightarrow{OQ} + \overrightarrow{OR}. \tag{2.1}$$

The vectors $\overrightarrow{OP}$, $\overrightarrow{OQ}$ and $\overrightarrow{OR}$ are collinear to the unit vectors **i**, **j**, **k**, respectively. Hence, there exist numbers x, y and z such that $\overrightarrow{OP} = x\mathbf{i}$,

$\overrightarrow{OQ} = y\mathbf{j}$ and $\overrightarrow{OR} = z\mathbf{k}$, and

$$\mathbf{a} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}. \tag{2.2}$$

Formula (2.2) is referred to as the *expansion* of **a** with respect to the unit vectors **i**, **j**, **k**.

Expansion (2.2) is applicable to any vector. Since **i**, **j**, **k** are of a unit length and any two of them are perpendicular to each other, we call the ordered triple (**i**, **j**, **k**) of unit vectors the *orthonormal basis.* It is easily seen that for any **a** the expansion is unique. The numbers x, y and z in (2.2) are called the *coordinates* of **a**. These are the same as the coordinates x, y and z of the terminal point A of **a**. So we write $\mathbf{a} = \{x, y, z\}$.

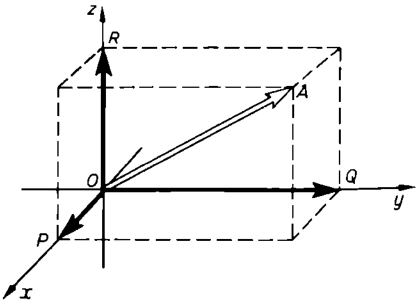


Fig. 2.20

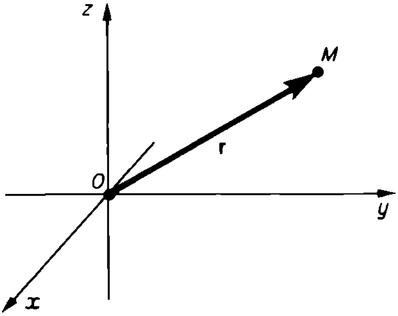


Fig. 2.21

Vectors $x\mathbf{i}$, $y\mathbf{j}$ and $z\mathbf{k}$ in (2.2) are called the *components* of **a**.

Let $\mathbf{a} = \{x_1, y_1, z_1\}$ and $\mathbf{b} = \{x_2, y_2, z_2\}$ be two vectors. Then **a** and **b** are equivalent if and only if their respective coordinates are equal, i.e.,

$$\mathbf{a} = \mathbf{b} \quad \text{iff} \quad \begin{cases} x_1 = x_2, \\ y_1 = y_2, \\ z_1 = z_2. \end{cases}$$

The *position vector* of a point $M(x, y, z)$ is the vector $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ which starts at the origin O of coordinates and ends at $M(x, y, z)$ (Fig. 2.21).

Linear operations on vectors specified by their coordinates. Let there be given two vectors $\mathbf{a} = \{x_1, y_1, z_1\}$ and $\mathbf{b} = \{x_2, y_2, z_2\}$ so that $\mathbf{a} = x_1\mathbf{i} + y_1\mathbf{j} + z_1\mathbf{k}$ and $\mathbf{b} = x_2\mathbf{i} + y_2\mathbf{j} + z_2\mathbf{k}$. Then, by laws of addition we get

$$\mathbf{a} + \mathbf{b} = (x_1\mathbf{i} + y_1\mathbf{j} + z_1\mathbf{k}) + (x_2\mathbf{i} + y_2\mathbf{j} + z_2\mathbf{k})$$
$$= (x_1 + x_2)\mathbf{i} + (y_1 + y_2)\mathbf{j} + (z_1 + z_2)\mathbf{k}$$

or

$$\mathbf{a} + \mathbf{b} = \{x_1 + x_2,\ y_1 + y_2,\ z_1 + z_2\}.$$

Hence, addition of vectors becomes addition of their corresponding coordinates.

Similarly, we have

$$\mathbf{a} - \mathbf{b} = \{x_1 - x_2,\ y_1 - y_2,\ z_1 - z_2\}.$$

Also

$$\lambda\mathbf{a} = \lambda x_1\mathbf{i} + \lambda y_1\mathbf{j} + \lambda z_1\mathbf{k}$$

or

$$\lambda\mathbf{a} = \{\lambda x_1,\ \lambda y_1,\ \lambda z_1\}.$$

Hence, multiplication of a vector by a scalar reduces to multiplication of all the coordinates of the vector by the scalar.

Let $\mathbf{a} = \{x_1, y_1, z_1\}$ and $\mathbf{b} = \{x_2, y_2, z_2\}$ be collinear vectors, and $\mathbf{b} \neq \mathbf{0}$. Then $\mathbf{a} = \mu\mathbf{b}$, i.e.,

$$x_1 = \mu x_2, \quad y_1 = \mu y_2, \quad z_1 = \mu z_2$$

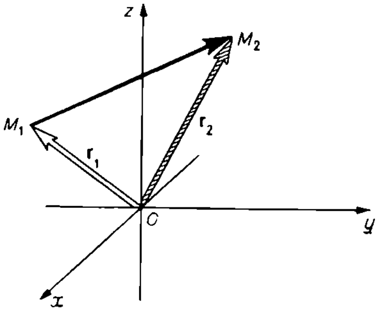


Fig. 2.22

or

$$\frac{x_1}{x_2} = \frac{y_1}{y_2} = \frac{z_1}{z_2}. \tag{2.3}$$

Conversely, if (2.3) is fulfilled, then $\mathbf{a} = \mu\mathbf{b}$, i.e., $\mathbf{a}$ and $\mathbf{b}$ are collinear.

In other words, the vectors $\mathbf{a}$ and $\mathbf{b}$ are collinear if and only if their coordinates are proportional.

Example. Find the coordinates of the vector $\overrightarrow{M_1M_2}$ where $M_1(x_1, y_1, z_1)$ is the initial point and $M_2(x_2, y_2, z_2)$ is the terminal point of the vector (Fig. 2.22).

◄ From Fig. 2.22 it follows that $\overrightarrow{M_1M_2} = \mathbf{r}_2 - \mathbf{r}_1$, where $\mathbf{r}_1$ and $\mathbf{r}_2$ are position vectors of M_1 and M_2, respectively. Thus

$$\overrightarrow{M_1M_2} = \{x_2 - x_1,\ y_2 - y_1,\ z_2 - z_1\}.$$

Hence, we conclude that the coordinates of $\overrightarrow{M_1M_2}$ are equal to the differences of the respective coordinates of M_2 and M_1. ►

2.4 Projection of a Vector onto an Axis

Let l be an axis and $\overrightarrow{AB}$ be a nonzero directed segment of l (Fig. 2.23).

The *magnitude* of the directed segment $\overrightarrow{AB}$ of the axis l is the number equal to the positive length of $\overrightarrow{AB}$ if the directions of $\overrightarrow{AB}$ and l coincide, and to the negative length of $\overrightarrow{AB}$ otherwise.

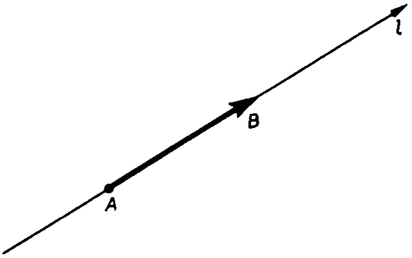


Fig. 2.23

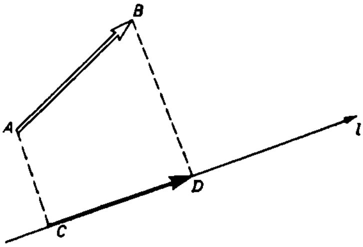


Fig. 2.24

Let $\overrightarrow{AB}$ be an arbitrary vector shown in Fig. 2.24. We drop from A and B perpendiculars onto the axis l. Thus we get the directed segment $\overrightarrow{CD}$ of l.

Definition. The *projection* of the vector $\overrightarrow{AB}$ onto the axis l is the magnitude of the directed segment $\overrightarrow{CD}$ as given above.

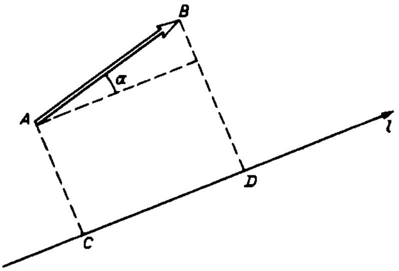


Fig. 2.25

Basic properties of projections. (1) The projection of a vector $\overrightarrow{AB}$ onto an axis l is equal to the product of the length of $\overrightarrow{AB}$ and the cosine of the angle between l and $\overrightarrow{AB}$ (Fig. 2.25), i.e.,

$$\text{proj}_l \overrightarrow{AB} = |\overrightarrow{AB}| \cos \alpha.$$

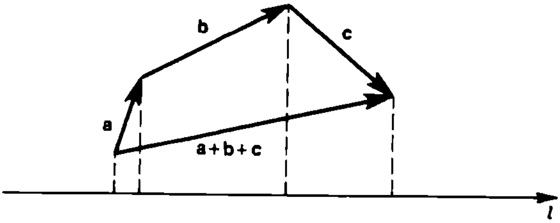


Fig. 2.26

(2) The projection of the sum of vectors onto an axis l is equal to the sum of the projections of the vectors onto the axis.

For example, from Fig. 2.26 if follows that

$$\text{proj}_l(\mathbf{a} + \mathbf{b} + \mathbf{c}) = \text{proj}_l \mathbf{a} + \text{proj}_l \mathbf{b} + \text{proj}_l \mathbf{c}.$$

2.5 Scalar Product of Two Vectors

Let there be given two nonzero vectors **a** and **b**.

Definition. A *scalar product* of two vectors **a** and **b** is a number $\mathbf{a} \cdot \mathbf{b}$ such that

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}|\, |\mathbf{b}| \cos \varphi = |\mathbf{a}|\, |\mathbf{b}| \cos (\widehat{\mathbf{a}, \mathbf{b}}), \tag{2.4}$$

where $\varphi = (\widehat{\mathbf{a}, \mathbf{b}})$ is the angle between **a** and **b**.

Observe that $|\mathbf{b}| \cos \varphi$ is the projection of **b** onto the axis determined by **a**. Then we may write (see Fig. 2.27)

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}|\, \text{proj}_a \mathbf{b}. \tag{2.5}$$

Similarly, we have

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{b}|\, \text{proj}_b \mathbf{a}.$$

Hence, a scalar product of two vectors is a number equal to the product of the length of one vector by the projection of the other vector onto the axis determined by the first vector.

When either **a** or **b** is a zero vector, we have $\mathbf{a} \cdot \mathbf{b} = 0$.

Properties of the scalar product. (1) The scalar product of $\mathbf{a}$ and $\mathbf{b}$ vanishes if and only if $\mathbf{a}$ and/or $\mathbf{b}$ are zero vectors, or if $\mathbf{a}$ and $\mathbf{b}$ are perpendicular (written $\mathbf{a} \perp \mathbf{b}$).

◄ This immediately follows from (2.4). ►

Since the direction of the zero vector is not defined we may set it perpendicular to any vector. So we may write the property in question as

$$\mathbf{a} \perp \mathbf{b} \quad \text{iff} \quad \mathbf{a}\cdot\mathbf{b} = 0.$$

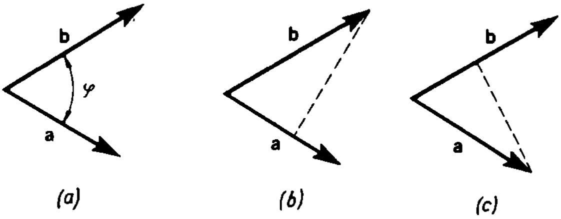


Fig. 2.27

(2) The scalar product obeys the commutative law, i.e.,

$$\mathbf{a}\cdot\mathbf{b} = \mathbf{b}\cdot\mathbf{a}.$$

◄ This immediately follows from (2.4), taking into account that $\cos\varphi$ is an even function. ►

(3) The scalar product obeys the distributive law relative to the sum of vectors, i.e.,

$$(\mathbf{a} + \mathbf{b})\cdot\mathbf{c} = \mathbf{a}\cdot\mathbf{c} + \mathbf{b}\cdot\mathbf{c}.$$

◄ Indeed,

$$\begin{aligned}(\mathbf{a} + \mathbf{b})\cdot\mathbf{c} &= |\mathbf{c}|\cdot\operatorname{proj}_{\mathbf{c}}(\mathbf{a} + \mathbf{b}) = |\mathbf{c}|\cdot(\operatorname{proj}_{\mathbf{c}}\mathbf{a} + \operatorname{proj}_{\mathbf{c}}\mathbf{b}) \\ &= |\mathbf{c}|\cdot\operatorname{proj}_{\mathbf{c}}\mathbf{a} + |\mathbf{c}|\cdot\operatorname{proj}_{\mathbf{c}}\mathbf{b} = \mathbf{a}\cdot\mathbf{c} + \mathbf{b}\cdot\mathbf{c}.\end{aligned}$$ ►

(4) If $\mathbf{a}$ or $\mathbf{b}$ is multiplied by a scalar λ, the scalar product of $\mathbf{a}$ and $\mathbf{b}$ is multiplied by λ, i.e.,

$$(\lambda\mathbf{a})\cdot\mathbf{b} = \mathbf{a}\cdot(\lambda\mathbf{b}) = \lambda(\mathbf{a}\cdot\mathbf{b}).$$

◄ Let $\lambda > 0$. Then

$$\lambda(\mathbf{a}\cdot\mathbf{b}) = \lambda|\mathbf{a}|\,|\mathbf{b}|\cos(\widehat{\mathbf{a}, \mathbf{b}})$$

and

$$(\lambda\mathbf{a})\cdot\mathbf{b} = |\lambda|\,|\mathbf{a}|\,|\mathbf{b}|\cos(\widehat{\lambda\mathbf{a}, \mathbf{b}}) = \lambda|\mathbf{a}|\,|\mathbf{b}|\cos(\widehat{\mathbf{a}, \mathbf{b}}),$$

since the angles $(\widehat{\mathbf{a}, \mathbf{b}})$ and $(\widehat{\lambda\mathbf{a}, \mathbf{b}})$ are equal (Fig. 2.28).

We proceed similarly when $\lambda < 0$.

The case of $\lambda = 0$ is trivial. ►

Remark. In general

$$(\mathbf{a}\cdot\mathbf{b})\cdot\mathbf{c} \neq \mathbf{a}\cdot(\mathbf{b}\cdot\mathbf{c}).$$

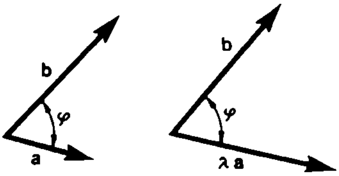


Fig. 2.28

Scalar product of two vectors specified by their coordinates. Let $\mathbf{a} = \{x_1, y_1, z_1\}$ and $\mathbf{b} = \{x_2, y_2, z_2\}$ be two vectors specified by their coordinates with respect to the orthonormal basis $(\mathbf{i}, \mathbf{j}, \mathbf{k})$.

We consider the scalar product of $\mathbf{a}$ and $\mathbf{b}$, i.e.,

$$\mathbf{a}\cdot\mathbf{b} = (x_1\mathbf{i} + y_1\mathbf{j} + z_1\mathbf{k})\cdot(x_2\mathbf{i} + y_2\mathbf{j} + z_2\mathbf{k}).$$

By the distributive law, we have

$$\mathbf{a}\cdot\mathbf{b} = x_1x_2(\mathbf{i}\cdot\mathbf{i}) + y_1x_2(\mathbf{j}\cdot\mathbf{i}) + z_1x_2(\mathbf{k}\cdot\mathbf{i}) + x_1y_2(\mathbf{i}\cdot\mathbf{j})$$
$$+ y_1y_2(\mathbf{j}\cdot\mathbf{j}) + z_1y_2(\mathbf{j}\cdot\mathbf{k}) + x_1z_2(\mathbf{i}\cdot\mathbf{k}) + y_1z_2(\mathbf{j}\cdot\mathbf{k}) + z_1z_2(\mathbf{k}\cdot\mathbf{k}).$$

Observe that

$$\mathbf{i}\cdot\mathbf{j} = \mathbf{i}\cdot\mathbf{k} = \mathbf{j}\cdot\mathbf{k} = 0,$$
$$\mathbf{i}\cdot\mathbf{i} = \mathbf{j}\cdot\mathbf{j} = \mathbf{k}\cdot\mathbf{k} = 1.$$

Whence we obtain

$$\mathbf{a}\cdot\mathbf{b} = x_1x_2 + y_1y_2 + z_1z_2. \tag{2.6}$$

We conclude that, given vectors $\mathbf{a}$ and $\mathbf{b}$ specified by their coordinates with respect to the orthonormal basis $(\mathbf{i}, \mathbf{j}, \mathbf{k})$, the scalar product of $\mathbf{a}$ and $\mathbf{b}$ is equal to the sum of the products of their respective coordinates.

Example. Compute the scalar product of

$$\mathbf{a} = 4\mathbf{i} - 2\mathbf{j} + \mathbf{k} \quad \text{and} \quad \mathbf{b} = 6\mathbf{i} + 3\mathbf{j} + 2\mathbf{k}.$$

◄ $\mathbf{a}\cdot\mathbf{b} = 4 \times 6 + (-2) \times 3 + 1 \times 2 = 20.$ ►

The *scalar square* of the vector $\mathbf{a}$ is $\mathbf{a}\cdot\mathbf{a} = a^2$.

Using (2.6) and setting **b** = **a**, we obtain

$$a^2 = x_1^2 + y_1^2 + z_1^2. \tag{2.7}$$

On the other hand,

$$a^2 = \mathbf{a}\cdot\mathbf{a} = |\mathbf{a}|^2\cdot\cos 0 = |\mathbf{a}|^2,$$

so that (2.7) gives

$$|\mathbf{a}| = \sqrt{x_1^2 + y_1^2 + z_1^2}\,. \tag{2.8}$$

Hence, the length of a vector specified by its coordinates with respect to the orthonormal basis is equal to the square root of the sum of squares of the coordinates.

Direction cosines. By definition, the scalar product of two vectors **a** and **b** is

$$\mathbf{a}\cdot\mathbf{b} = |\mathbf{a}|\,|\mathbf{b}|\cos\varphi,$$

where φ is the angle between **a** and **b**.

Provided that both **a** and **b** are nonzero vectors, the angle between **a** and **b** is given by

$$\cos\varphi = \frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}|\,|\mathbf{b}|}. \tag{2.9}$$

For $\mathbf{a} = \{x_1, y_1, z_1\}$ and $\mathbf{b} = \{x_2, y_2, z_2\}$ (2.9) becomes

$$\cos\varphi = \frac{x_1x_2 + y_1y_2 + z_1z_2}{\sqrt{x_1^2 + y_1^2 + z_1^2}\cdot\sqrt{x_2^2 + y_2^2 + z_2^2}}. \tag{2.10}$$

Example. Find the angle between $\mathbf{a} = \{2, -4, 4\}$ and $\mathbf{b} = \{-3, 2, 6\}$.

◄ Using (2.10), we get

$$\cos\varphi = \frac{-6 - 8 + 24}{\sqrt{4 + 16 + 16}\cdot\sqrt{9 + 4 + 36}} = \frac{5}{21}.$$

The value of φ is easily obtained from tables of trigonometric functions. ►

Let **b** = **i**, that is, $\mathbf{b} = \{1, 0, 0\}$. Then for any $\mathbf{a} = \{x_1, y_1, z_1\} \neq 0$ and the unit vector **i** we have

$$\cos\alpha = \frac{\mathbf{a}\cdot\mathbf{i}}{|\mathbf{a}|}$$

or

$$\cos\alpha = \frac{x_1}{\sqrt{x_1^2 + y_1^2 + z_1^2}}, \tag{2.11}$$

where α is the angle between **a** and the x-axis.

Similarly,

$$\cos\beta = \frac{y_1}{\sqrt{x_1^2 + y_1^2 + z_1^2}} = \frac{\mathbf{a}\cdot\mathbf{j}}{|\mathbf{a}|}, \tag{2.12}$$

$$\cos\gamma = \frac{z_1}{\sqrt{x_1^2 + y_1^2 + z_1^2}} = \frac{\mathbf{a}\cdot\mathbf{k}}{|\mathbf{a}|}. \tag{2.13}$$

Formulas (2.11)-(2.13) define the *direction cosines* of the vector **a**, i.e., the cosines of angles between **a** and the coordinate axes (Fig. 2.29).

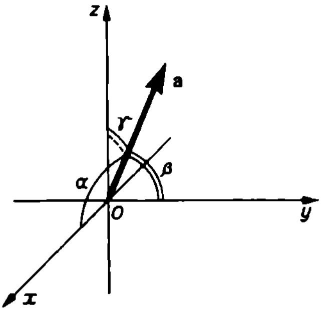


Fig. 2.29

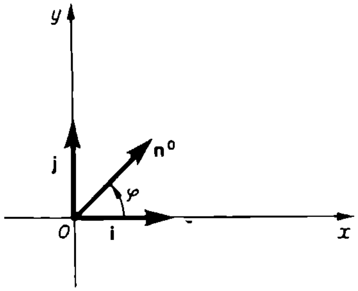


Fig. 2.30

Example. Find the coordinates of the unit vector $\mathbf{n}^0$.

◄ Let $\mathbf{n}^0 = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$. Since $|\mathbf{n}^0| = 1$, we get

$$\mathbf{n}^0\cdot\mathbf{i} = x = |\mathbf{n}^0|\,|\mathbf{i}|\cos(\widehat{\mathbf{n}^0, \mathbf{i}}) = \cos\alpha,$$

$$\mathbf{n}^0\cdot\mathbf{j} = y = \cos\beta,$$

$$\mathbf{n}^0\cdot\mathbf{k} = z = \cos\gamma.$$

Thus, the coordinates of the unit vector are direction cosines of this vector, i.e., cosines of the angles between $\mathbf{n}^0$ and the coordinate axes, so that

$$\mathbf{n}^0 = \mathbf{i}\cos\alpha + \mathbf{j}\cos\beta + \mathbf{k}\cos\gamma.$$

Whence

$$(\mathbf{n}^0)^2 = \mathbf{n}^0\cdot\mathbf{n}^0 = 1 = \cos^2\alpha + \cos^2\beta + \cos^2\gamma.$$ ►

Example. Let $\mathbf{n}^0$ be orthogonal to the z-axis, i.e.,

$$\mathbf{n}^0 = x\mathbf{i} + y\mathbf{j}.$$

Then the coordinates x and y of $\mathbf{n}^0$ are

$$x = \cos\varphi \quad \text{and} \quad y = \sin\varphi,$$

so that (Fig. 2.30)

$$\mathbf{n}^0 = \mathbf{i}\cos\varphi + \mathbf{j}\sin\varphi.$$

2.6 Vector Product of Two Vectors

Definition. The *vector product* of two noncollinear vectors **a** and **b** is a vector $\mathbf{a} \times \mathbf{b}$ such that

(1) the length of $\mathbf{a} \times \mathbf{b}$ is equal to $|\mathbf{a}|\,|\mathbf{b}|\sin\varphi$, where φ is the angle between **a** and **b** (Fig. 2.31);

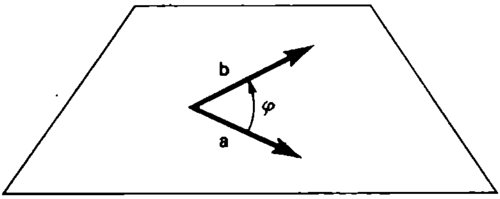


Fig. 2.31

(2) $\mathbf{a} \times \mathbf{b}$ is perpendicular to both **a** and **b**, i.e., $\mathbf{a} \times \mathbf{b}$ is perpendicular to the plane determined by **a** and **b**;

(3) $\mathbf{a} \times \mathbf{b}$ is so directed that the shortest rotation from **a** to **b** is made counterclockwise as seen from the terminal point of the vector $\mathbf{a} \times \mathbf{b}$ (Fig. 2.32). In other words, the ordered triple of vectors $(\mathbf{a}, \mathbf{b}, \mathbf{a} \times \mathbf{b})$ is *right-handed*, i.e., the thumb, the index finger and the middle finger of the right hand point along the directions of **a**, **b** and $\mathbf{a} \times \mathbf{b}$, respectively.

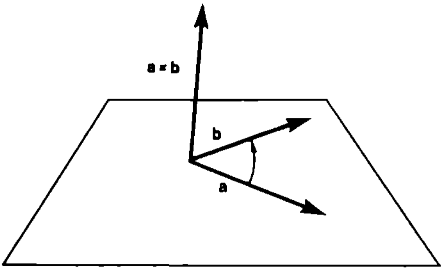


Fig. 2.32

When **a** and **b** are collinear, $\mathbf{a} \times \mathbf{b} = \mathbf{0}$.

The length of the vector product of **a** and **b** denoted by

$$|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}|\,|\mathbf{b}| \sin \varphi \tag{2.14}$$

is numerically equal to the area $S_{\square}$ of the parallelogram with sides **a** and **b** (Fig. 2.33), i.e.,

$$|\mathbf{a} \times \mathbf{b}| = S_{\square}.$$

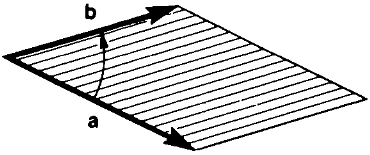


Fig. 2.33

Properties of the vector product. (1) The vector product of **a** and **b** is equal to the zero vector if and only if **a** and/or **b** are zero vectors, or if **a** and **b** are collinear. (Recall that if **a** and **b** are collinear then the angle between **a** and **b** is either zero or π.)

◄ This immediately follows from $|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}|\,|\mathbf{b}| \sin \varphi$. ►

We may set the zero vector collinear to any vector and express the property as

$$\mathbf{a} \parallel \mathbf{b} \quad \text{iff} \quad \mathbf{a} \times \mathbf{b} = \mathbf{0}.$$

(2) The vector product obeys the anticommutative law, i.e.,

$$\mathbf{b} \times \mathbf{a} = -(\mathbf{a} \times \mathbf{b}). \tag{2.15}$$

◄ Indeed, $\mathbf{a} \times \mathbf{b}$ and $\mathbf{b} \times \mathbf{a}$ are collinear and of the same length. Since the shortest rotation from **a** to **b** is counterclockwise for $\mathbf{a} \times \mathbf{b}$ and is clockwise for $\mathbf{b} \times \mathbf{a}$, we conclude that $\mathbf{a} \times \mathbf{b}$ and $\mathbf{b} \times \mathbf{a}$ are of opposite directions (Fig. 2.34) ►.

(3) The vector product obeys the distributive law relative to the sum of vectors, i.e.,

$$(\mathbf{a} + \mathbf{b}) \times \mathbf{c} = \mathbf{a} \times \mathbf{c} + \mathbf{b} \times \mathbf{c}.$$

(4) If **a** or **b** is multiplied by a scalar λ the vector product of **a** and **b** is multiplied by λ, i.e.,

$$(\lambda \mathbf{a}) \times \mathbf{b} = \mathbf{a} \times (\lambda \mathbf{b}) = \lambda(\mathbf{a} \times \mathbf{b}).$$

Remark. The vector product is not associative. Indeed, the identity $(\mathbf{a} \times \mathbf{b}) \times \mathbf{c} = \mathbf{a} \times (\mathbf{b} \times \mathbf{c})$ is not generally true. For example,

$$(\mathbf{i} \times \mathbf{j}) \times \mathbf{j} = -\mathbf{i} \quad \text{but} \quad \mathbf{i} \times (\mathbf{j} \times \mathbf{j}) = \mathbf{0}.$$

Vector product of two vectors specified by their coordinates. Let $\mathbf{a} = \{x_1, y_1, z_1\}$ and $\mathbf{b} = \{x_2, y_2, z_2\}$ be two vectors specified by their coordinates with respect to the orthonormal basis $(\mathbf{i}, \mathbf{j}, \mathbf{k})$.

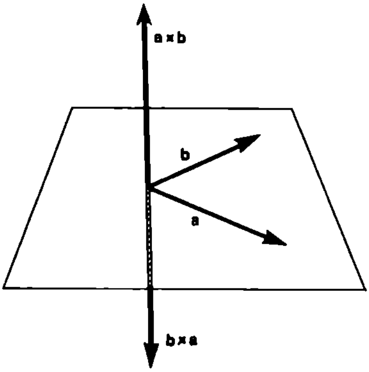


Fig. 2.34

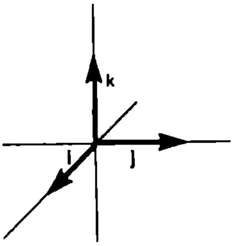


Fig. 2.35

By the distributive law, we obtain

$$\begin{aligned}\mathbf{a} \times \mathbf{b} &= (x_1\mathbf{i} + y_1\mathbf{j} + z_1\mathbf{k}) \times (x_2\mathbf{i} + y_2\mathbf{j} + z_2\mathbf{k}) \\ &= x_1x_2(\mathbf{i} \times \mathbf{i}) + x_1y_2(\mathbf{i} \times \mathbf{j}) + x_1z_2(\mathbf{i} \times \mathbf{k}) + y_1x_2(\mathbf{j} \times \mathbf{i}) \\ &+ y_1y_2(\mathbf{j} \times \mathbf{j}) + y_1z_2(\mathbf{j} \times \mathbf{k}) + z_1x_2(\mathbf{k} \times \mathbf{i}) + z_1y_2(\mathbf{k} \times \mathbf{j}) \\ &+ z_1z_2(\mathbf{k} \times \mathbf{k}). \end{aligned} \tag{2.16}$$

The vector products of basis vectors (see Fig. 2.35) are

$$\mathbf{i} \times \mathbf{i} = \mathbf{j} \times \mathbf{j} = \mathbf{k} \times \mathbf{k} = \mathbf{0},$$

$$\mathbf{i} \times \mathbf{j} = \mathbf{k}, \quad \mathbf{j} \times \mathbf{k} = \mathbf{i}, \quad \mathbf{k} \times \mathbf{i} = \mathbf{j},$$

$$\mathbf{j} \times \mathbf{i} = -\mathbf{k}, \quad \mathbf{k} \times \mathbf{j} = -\mathbf{i}, \quad \mathbf{i} \times \mathbf{k} = -\mathbf{j}.$$

Substituting these relations into (2.16), we obtain

$$\begin{aligned}\mathbf{a} \times \mathbf{b} &= -x_2y_1\mathbf{k} + x_2z_1\mathbf{j} + x_1y_2\mathbf{k} - z_1y_2\mathbf{i} + y_1z_2\mathbf{i} - x_1z_2\mathbf{j} \\ &= \mathbf{i}(y_1z_2 - y_2z_1) + \mathbf{j}(x_2z_1 - x_1z_2) + \mathbf{k}(x_1y_2 - x_2y_1). \end{aligned} \tag{2.17}$$

To memorize (2.17) it is helpful to rewrite it in the from of the third-order determinant as

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{vmatrix}. \tag{2.18}$$

Expanding (2.18) with respect to the first row, we obtain (2.17).

Examples. (1) Compute the area of a parallelogram with sides $\mathbf{a} = \mathbf{i} + \mathbf{j} - \mathbf{k}$ and $\mathbf{b} = 2\mathbf{i} + \mathbf{j} - \mathbf{k}$.

◄ The area we are seeking for is $S_{\square} = |\mathbf{a} \times \mathbf{b}|$. By (2.18), we get

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 1 & -1 \\ 2 & 1 & -1 \end{vmatrix} = \mathbf{i}\cdot 0 - \mathbf{j}\cdot 1 + \mathbf{k}\cdot(-1) = -\mathbf{j} + \mathbf{k}.$$

Whence

$$S_{\square} = \sqrt{1 + 1} = \sqrt{2}. \quad ►$$

(2) Compute the area of the triangle OAB shown in Fig. 2.36.

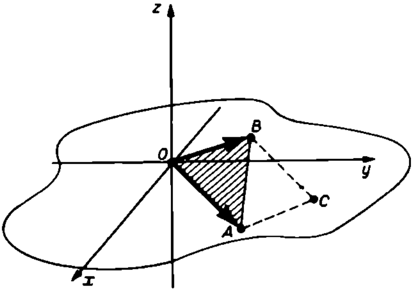


Fig. 2.36

◄ We clearly see that the area $S_{\triangle}$ of OAB is half the area $S_{\square}$ of the parallelogram $OACB$. Computing the vector product of $\mathbf{a} = \overrightarrow{OA}$ and $\mathbf{b} = \overrightarrow{OB}$, we obtain

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x_1 & y_1 & 0 \\ x_1 & y_2 & 0 \end{vmatrix} = (x_1y_2 - x_2y_1)|\mathbf{k}.$$

Whence

$$S_{\square} = |x_1y_2 - x_2y_1| \quad \text{and} \quad S_{\triangle} = \frac{1}{2}|x_1y_2 - x_2y_1|. \quad ►$$

2.7 Mixed Products of Three Vectors

Scalar triple product. Let there be given three vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$. We find the vector product $\mathbf{a} \times \mathbf{b}$ of $\mathbf{a}$ and $\mathbf{b}$, and then the scalar product of $\mathbf{a} \times \mathbf{b}$ and $\mathbf{c}$, i.e., the number $(\mathbf{a} \times \mathbf{b})\cdot\mathbf{c}$.

The number

$$(\mathbf{a} \times \mathbf{b})\cdot\mathbf{c}$$

is called the *scalar triple product* of three vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$.

Geometric interpretation of the scalar triple product. Let three vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ be applied to some point O as shown in Fig. 2.37. If the points O, A, B and C lie in the same plane, i.e., if the vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are *coplanar*, their scalar triple product $(\mathbf{a} \times \mathbf{b})\cdot\mathbf{c}$ is equal to zero, i.e., $(\mathbf{a} \times \mathbf{b})\cdot\mathbf{c} = 0$, since the vector $\mathbf{a} \times \mathbf{b}$ is perpendicular to the plane containing the vectors $\mathbf{a}$ and $\mathbf{b}$ and hence $\mathbf{c}$. In other words, the vectors $\mathbf{a} \times \mathbf{b}$ and $\mathbf{c}$ are perpendicular to each other and their scalar product is equal to zero.

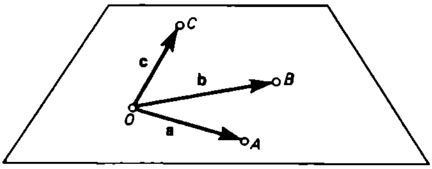


Fig. 2.37

When the points O, A, B, C do not lie in the same plane, i.e., when $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are noncoplanar, we construct a parallelepiped with edges OA, OB and OC (Fig. 2.38).

By definition

$$\mathbf{a} \times \mathbf{b} = S\mathbf{e},$$

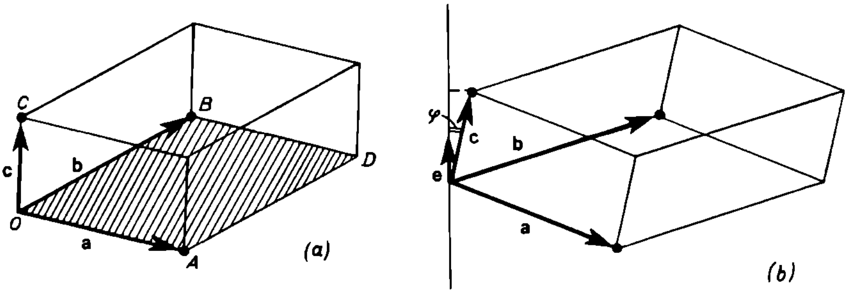


Fig. 2.38

where S is the area of the parallelogram $OADB$, $\mathbf{e}$ is the unit vector perpendicular to both $\mathbf{a}$ and $\mathbf{b}$ so that the ordered triple of vectors $(\mathbf{a}, \mathbf{b}, \mathbf{e})$ is right-handed, i.e., the thumb, the index finger and the middle finger of the right hand point along the directions of $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{e}$, respectively.

Computing the scalar product of $\mathbf{a} \times \mathbf{b}$ and $\mathbf{c}$, we obtain

$$(\mathbf{a} \times \mathbf{b})\cdot\mathbf{c} = (S\mathbf{e})\cdot\mathbf{c} = S(\mathbf{e}\cdot\mathbf{c}) = S\,\mathrm{proj}_e\,\mathbf{c}.$$

The number $S\,\mathrm{proj}_e\,\mathbf{c}$ is equal to the positive height h of the parallelepiped if the angle φ between $\mathbf{e}$ and $\mathbf{c}$ is acute, i.e., if the ordered triple of vectors $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ is right-handed, and to the negative height if the angle φ is obtuse, i.e., if $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ is the left-handed ordered triple of vectors, so that

$$(\mathbf{a} \times \mathbf{b})\cdot\mathbf{c} = \pm Sh = \pm V.$$

Therefore the scalar triple product of $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ is numerically equal to the volume V of the parallelepiped with edges $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ when the triple $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ is right-handed, and to minus V when the triple $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ is left-handed.

The geometry reveals that multiplying the three vectors in any other order, we obtain either $+V$ or $-V$.

Notice that when the triple $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ is right-handed, both $(\mathbf{b}, \mathbf{c}, \mathbf{a})$ and $(\mathbf{c}, \mathbf{a}, \mathbf{b})$ are also right-handed, while all the triples $(\mathbf{b}, \mathbf{a}, \mathbf{c})$, $(\mathbf{a}, \mathbf{c}, \mathbf{b})$ and $(\mathbf{c}, \mathbf{b}, \mathbf{a})$ are left-handed. Thus

$$\begin{aligned}(\mathbf{a} \times \mathbf{b})\cdot\mathbf{c} &= (\mathbf{b} \times \mathbf{c})\cdot\mathbf{a} = (\mathbf{c} \times \mathbf{a})\cdot\mathbf{b}\\ &= -(\mathbf{b} \times \mathbf{a})\cdot\mathbf{c} = -(\mathbf{a} \times \mathbf{c})\cdot\mathbf{b} = -(\mathbf{c} \times \mathbf{b})\cdot\mathbf{a}.\end{aligned}$$

Once again we shall emphasize the following: the scalar triple product of three vectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ is equal to zero if and only if $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are coplanar.

In symbols,

$$\mathbf{a}, \mathbf{b}, \mathbf{c} \text{ are coplanar iff } (\mathbf{a} \times \mathbf{b})\cdot\mathbf{c} = 0.$$

Scalar triple product of three vectors specified by their coordinates. Let $\mathbf{a} = \{x_1, y_1, z_1\}$, $\mathbf{b} = \{x_2, y_2, z_2\}$ and $\mathbf{c} = \{x_3, y_3, z_3\}$ be three vectors specified by their coordinates with respect to the orthonormal basis $(\mathbf{i}, \mathbf{j}, \mathbf{k})$. Let us express the scalar triple product $(\mathbf{a} \times \mathbf{b})\cdot\mathbf{c}$ in terms of coordinates of $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$.

We have

$$\begin{aligned}\mathbf{a} \times \mathbf{b} &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{vmatrix}\\ &= \mathbf{i}(y_1z_2 - y_2z_1) + \mathbf{j}(x_2z_1 - x_1z_2) + \mathbf{k}(x_1y_2 - x_2y_1).\end{aligned}$$

Whence

$$(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = x_3 \begin{vmatrix} y_1 & z_1 \\ y_2 & z_2 \end{vmatrix} - y_3 \begin{vmatrix} x_1 & z_1 \\ x_2 & z_2 \end{vmatrix} + z_3 \begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix} = \begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{vmatrix}.$$

Thus the scalar triple product of three vectors specified by their coordinates with respect to the orthonormal basis (**i**, **j**, **k**) is equal to the third-order determinant whose rows are coordinates of the vectors being multiplied.

The necessary and sufficient condition for three vectors $\mathbf{a} = \{x_1, y_1, z_1\}$, $\mathbf{b} = \{x_2, y_2, z_2\}$ and $\mathbf{c} = \{x_3, y_3, z_3\}$ to be coplanar may be written as

$$\begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{vmatrix} = 0.$$

Example. Verify whether the vectors $\mathbf{a} = \{7, 4, 6\}$, $\mathbf{b} = \{2, 1, 1\}$ and $\mathbf{c} = \{19, 11, 17\}$ are coplanar.

◄ The vectors are coplanar if

$$\Delta = \begin{vmatrix} 7 & 4 & 6 \\ 2 & 1 & 1 \\ 19 & 11 & 17 \end{vmatrix}$$

is equal to zero, and noncoplanar otherwise.

Expanding Δ with respect to the first row, we obtain

$$\Delta = 7 \times 6 - 4 \times 15 + 6 \times 3 = 0.$$

Thus we conclude that the vectors are coplanar. ►

Vector triple product. The vector triple product $\mathbf{a} \times (\mathbf{b} \times \mathbf{c})$ of three vectors **a**, **b** and **c** is the vector perpendicular to both **a** and $\mathbf{b} \times \mathbf{c}$. Hence the vector $\mathbf{a} \times (\mathbf{b} \times \mathbf{c})$ lies in the plane determined by **b** and **c** and may be expanded with respect to **b** and **c**. It is easy to show that

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{b} \cdot (\mathbf{a} \cdot \mathbf{c}) - \mathbf{c} \cdot (\mathbf{a} \cdot \mathbf{b}).$$

Exercises

1. The vectors $\overrightarrow{AB} = \mathbf{c}$, $\overrightarrow{BC} = \mathbf{a}$ and $\overrightarrow{CA} = \mathbf{b}$ are sides of a triangle. Express the medians of the triangle in terms of **a**, **b** and **c**.
2. Let **p**, **q** and **p** + **q** be three vectors applied to a common point. Find the relationship between **p** and **q** provided that **p** + **q** bisects the angle between **p** and **q**.

3. Compute the lengths of the diagonals of a parallelogram with sides $\mathbf{a} = 5\mathbf{p} + 2\mathbf{q}$ and $\mathbf{b} = \mathbf{p} - 3\mathbf{q}$ provided that $|\mathbf{p}| = 2\sqrt{2}$, $|\mathbf{q}| = 3$ and $(\widehat{\mathbf{p}, \mathbf{q}}) = \pi/4$.

4. Prove that the diagonals of the rhombus are perpendicular to each other.

5. Compute the scalar product of $\mathbf{a} = 4\mathbf{i} + 7\mathbf{j} + 3\mathbf{k}$ and $\mathbf{b} = 3\mathbf{i} - 5\mathbf{j} + \mathbf{k}$.

6. Find the unit vector $\mathbf{a}^0$ which is parallel to $\mathbf{a} = \{6, 7, -6\}$.

7. Find the projection of the vector $\mathbf{a} = \mathbf{i} + \mathbf{j} - \mathbf{k}$ onto the axis determined by the vector $\mathbf{b} = 2\mathbf{i} - \mathbf{j} - 3\mathbf{k}$.

8. Find the cosine of the angle between the vectors $\overrightarrow{AB}$ and $\overrightarrow{AC}$, given $A(-4, 0, 4)$, $B(-1, 6, 7)$ and $C(1, 10, 9)$.

9. Find the unit vector $\mathbf{p}^0$ which is perpendicular both to the vector $\mathbf{a} = \{3, 6, 8\}$ and to the x-axis.

10. Find the sine of the angle between the diagonals of the parallelogram with sides $\mathbf{a} = 2\mathbf{i} + \mathbf{j} - \mathbf{k}$ and $\mathbf{b} = \mathbf{i} - 3\mathbf{j} + \mathbf{k}$.

11. Find the height h of the parallelepiped with edges $\mathbf{a} = 3\mathbf{i} + 2\mathbf{j} - 5\mathbf{k}$, $\mathbf{b} = \mathbf{i} - \mathbf{j} + 4\mathbf{k}$ and $\mathbf{c} = \mathbf{i} - 3\mathbf{j} + \mathbf{k}$ provided that the base of the parallelepiped is the parallelogram with sides $\mathbf{a}$ and $\mathbf{b}$.

Answers

1. $\overrightarrow{AM} = \mathbf{c} + \frac{\mathbf{a}}{2}$ or $\overrightarrow{AM} = \frac{\mathbf{c} - \mathbf{b}}{2}$; $\overrightarrow{BN} = \mathbf{a} + \frac{\mathbf{b}}{2}$ or $\overrightarrow{BN} = \frac{\mathbf{a} - \mathbf{c}}{2}$; $\overrightarrow{CP} = \mathbf{b} + \frac{\mathbf{c}}{2}$ or $\overrightarrow{CP} = \frac{\mathbf{b} - \mathbf{a}}{2}$. **2.** $|\mathbf{p}| = |\mathbf{q}|$ since the diagonal of the parallelogram bisects the angle only if the parallelogram is the rhombus. **3.** $|\mathbf{a} + \mathbf{b}| = 15$, $|\mathbf{a} - \mathbf{b}| = \sqrt{593}$. **5.** -20. **6.** $\mathbf{a}^0 = \left\{\frac{6}{11}, \frac{7}{11}, -\frac{6}{11}\right\}$ or $\mathbf{a}^0 = \left\{-\frac{6}{11}, -\frac{7}{11}, \frac{6}{11}\right\}$. **7.** $\mathrm{proj}_b\, \mathbf{a} = \frac{4}{\sqrt{14}}$. **8.** $\cos\varphi = 1$, $\varphi = 0$. **9.** $\mathbf{p}^0 = \pm\left\{0, -\frac{4}{5}, \frac{3}{5}\right\}$. **10.** $\sin\varphi = \sqrt{\frac{248}{273}}$. **11.** $h = \frac{49}{\sqrt{323}}$.

Chapter 3

The Line and the Plane

3.1 The Plane

Normal and general equations of a plane. We may determine a plane in space by specifying, relative to a given point O, two quantities, namely (1) the perpendicular distance p between O and the plane, i.e., the length of the perpendicular OT dropped from O onto the plane, and (2) the unit vector $\mathbf{n}^0$ which starts at O and is perpendicular to the plane, as shown in Fig. 3.1.

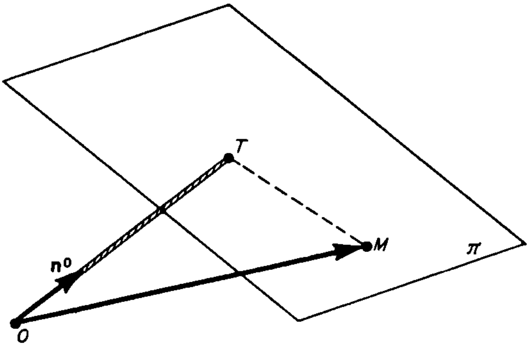


Fig. 3.1

Suppose that an arbitrary point M moves in the plane π. Whatever the path of moving, the position vector $\mathbf{r} = \overrightarrow{OM}$ varies so that the relation

$$\text{proj}_{n^0} \overrightarrow{OM} = p \tag{3.1}$$

remains valid unless M leaves the plane. In other words, relation (3.1) describes a property peculiar to all points of the plane, i.e., relation (3.1) is the equation of a plane. Observing that $\text{proj}_{n^0} \overrightarrow{OM} = \mathbf{r}\cdot\mathbf{n}^0$, we may write (3.1) as

$$\mathbf{r}\cdot\mathbf{n}^0 - p = 0. \tag{3.2}$$

Equation (3.2) is called the *normal vector equation of a plane.* The position vector $\mathbf{r}$ is sometimes referred to as a *moving position vector.*

Let us set up in space a Cartesian coordinate system with the origin at O. Then $\mathbf{n}^0 = \{\cos\alpha, \cos\beta, \cos\gamma\}$ and $\mathbf{r} = \{x, y, z\}$, and the equation (3.2) takes the form

$$x\cos\alpha + y\cos\beta + z\cos\gamma - p = 0, \tag{3.3}$$

in which case

(1) $\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1$,

(2) the term $(-p)$ is nonpositive.

Equation (3.3) is called the *normal coordinate equation of a plane.*

Since (3.3) is the first-degree equation in x, y and z, we conclude that in three-dimensional space a plane is determined by the first-degree equation in Cartesian coordinates x, y and z.

The converse is also true, namely any first-degree equation in Cartesian coordinates x, y and z defines a plane.

◄ Indeed, let us consider a general first-degree equation

$$Ax + By + Cz + D = 0 \quad (A^2 + B^2 + C^2 > 0). \tag{3.4}$$

Multiplying (3.4) by a scalar μ, we obtain

$$\mu Ax + \mu By + \mu Cz + \mu D = 0. \tag{3.5}$$

We choose μ such that (3.5) reduces to the normal equation (3.3). It suffices to set

$$\mu A = \cos\alpha, \quad \mu B = \cos\beta, \quad \mu C = \cos\gamma, \quad \mu D = -p. \tag{3.6}$$

From (3.6) we obtain

$$\mu^2(A^2 + B^2 + C^2) = 1.$$

Whence

$$\mu = \pm\frac{1}{\sqrt{A^2 + B^2 + C^2}}. \tag{3.7}$$

From $\mu D = -p \leqslant 0$ it follows that the sign in (3.7) must be opposite to the sign of D. Substituting (3.7) into (3.6), we find expressions for $\cos\alpha$, $\cos\beta$, $\cos\gamma$ and p in terms of the coefficients of equation (3.4). In other words we reduce equation (3.4) to the normal coordinate equation (3.3). The scalar μ is said to be a *normalizing factor.* Since the normal coordinate equation defines a plane, so does equation (3.4), called the *general equation of a plane* in three-dimensional space. ►

Therefore, any first-degree equation in x, y and z defines a plane in space as a set of points whose coordinates satisfy this equation.

We shall call any nonzero vector perpendicular to a given plane the *nor-*

mal vector to this plane. This implies that the vector $\mathbf{n} = \{A, B, C\}$ is a vector normal to the plane (3.4). Hence, A, B and C are easy to interpret as coordinates of the vector $\mathbf{n} = \{A, B, C\}$ normal to the plane (3.4) (see Fig. 3.2). Any other normal vector is obtained by multiplying $\mathbf{n}$ by a nonzero scalar.

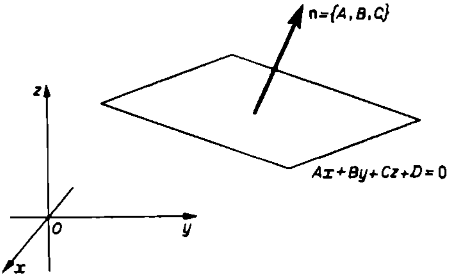


Fig. 3.2

Equation of a plane which passes through a given point and is perpendicular to a given direction. We wish to find an equation of a plane which passes through a point M_0 specified by the position vector $\mathbf{r}_0 = \{x_0, y_0, z_0\}$ and is perpendicular to the vector $\mathbf{n} = \{A, B, C\}$.

Let $\mathbf{r} = \{x, y, z\}$ be a position vector of an arbitrary point M of a plane (Fig. 3.3). The vector $\overrightarrow{M_0M} = \mathbf{r} - \mathbf{r}_0$ lies in the plane and, consequently, is perpendicular to the vector $\mathbf{n}$. So the scalar product of the vectors $\mathbf{r} - \mathbf{r}_0$ and $\mathbf{n}$ is equal to zero, i.e.,

$$(\mathbf{r} - \mathbf{r}_0)\cdot\mathbf{n} = 0. \tag{3.8}$$

The identity (3.8) remains valid for all points of the plane and does not hold for any point outside the plane. Hence, (3.8) is the normal vector equation we are seeking for. Expressing (3.8) in terms of coordinates of

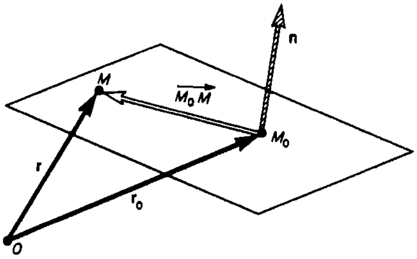


Fig. 3.3

$\mathbf{r} - \mathbf{r}_0$ and $\mathbf{n}$, we obtain the desired equation

$$A(x - x_0) + B(y - y_0) + C(z - z_0) = 0. \tag{3.9}$$

The angle between two planes. We consider two planes specified by the equations

$$A_1x + B_1y + C_1z + D_1 = 0 \quad (A_1^2 + B_1^2 + C_1^2 > 0),$$
$$A_2x + B_2y + C_2z + D_2 = 0 \quad (A_2^2 + B_2^2 + C_2^2 > 0).$$

The *angle* between two planes is any of the two dihedral angles formed by these planes (Fig. 3.4). When two planes are parallel the angle between them is equal either to zero or to π.

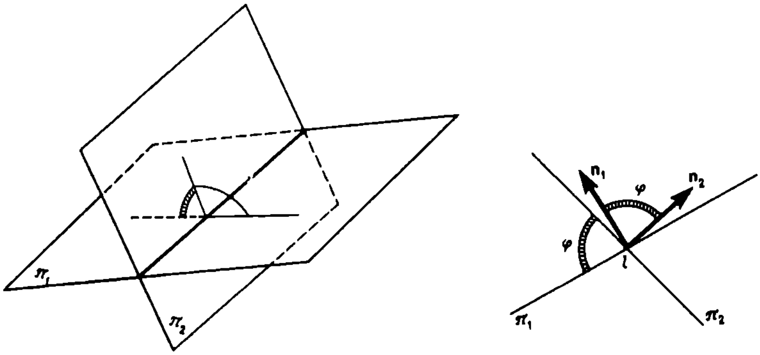


Fig. 3.4

We easily see from Fig. 3.4 that one of the two dihedral angles is equal to the angle φ between the vectors $\mathbf{n}_1 = \{A_1, B_1, C_1\}$ and $\mathbf{n}_2 = \{A_2, B_2, C_2\}$, whence we have

$$\cos\varphi = \frac{\mathbf{n}_1 \cdot \mathbf{n}_2}{|\mathbf{n}_1| \cdot |\mathbf{n}_2|} = \frac{A_1A_2 + B_1B_2 + C_1C_2}{\sqrt{A_1^2 + B_1^2 + C_1^2} \cdot \sqrt{A_2^2 + B_2^2 + C_2^2}}. \tag{3.10}$$

Conditions of perpendicularity and parallelism of two planes. If two planes are perpendicular to each other, their respective normal vectors are also perpendicular. Hence, there holds

$$\mathbf{n}_1 \cdot \mathbf{n}_2 = 0$$

or

$$A_1A_2 + B_1B_2 + C_1C_2 = 0.$$

This is the *condition of perpendicularity* of two planes.

If two planes are parallel, their respective normal vectors are collinear, i.e., $\mathbf{n}_1 = \lambda\mathbf{n}_2$. Expressing this identity in terms of coordinates of $\mathbf{n}_1$ and $\mathbf{n}_2$, we obtain

$$A_1 = \lambda A_2, \quad B_1 = \lambda B_2, \quad C_1 = \lambda C_2$$

or

$$\frac{A_1}{A_2} = \frac{B_1}{B_2} = \frac{C_1}{C_2}.$$

This is the *condition of parallelism of two planes.*

Problem. Find the equation of the plane passing through three noncollinear points $M_1(x_1, y_1, z_1)$, $M_2(x_2, y_2, z_2)$ and $M_3(x_3, y_3, z_3)$.

3.2 Straight Line in a Plane

An approach identical to that we used in preceding sections is fully applicable when we consider a (straight) line in a plane.

Here we present the major results with reference to accompanying figures.

1. *Standard forms of the equation of a line:*

(a) The normal vector equation (Fig. 3.5)

$$\mathbf{r} \cdot \mathbf{n}^0 - p = 0.$$

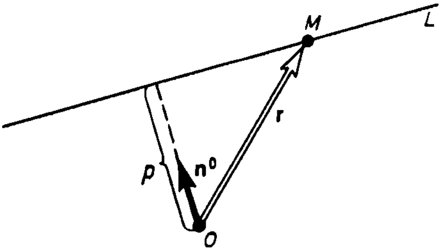


Fig. 3.5

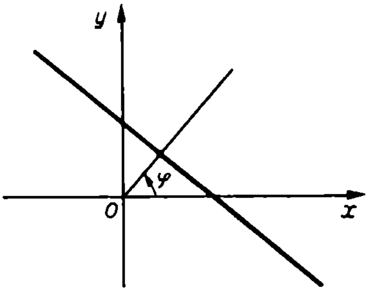


Fig. 3.6

(b) The normal coordinate equation (Fig. 3.6)

$$x \cos\varphi + y \sin\varphi - p = 0.$$

(c) The general equation (Fig. 3.7)

$$Ax + By + C = 0 \quad (A^2 + B^2 > 0),$$

where A and B are coordinates of the vector $\mathbf{n} = \{A, B\}$ normal to the line.

Multiplying the general equation by

$$\mu = \pm \frac{1}{\sqrt{A^2 + B^2}}$$

we may reduce the general equation to the normal coordinate equation.

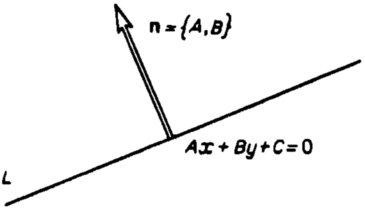


Fig. 3.7

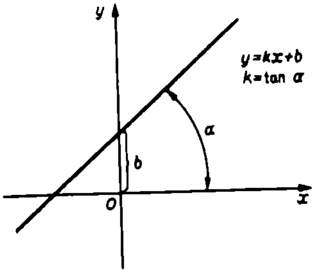


Fig. 3.8

The sign of μ is chosen so that

$$\mu C = -p \leqslant 0.$$

We mention two interesting cases:

(a) If $B \neq 0$, then setting $k = -\frac{A}{B}$ and $b = -\frac{C}{B}$ (see Fig. 3.8), we obtain the *slope intercept form* of the equation of a line

$$y = kx + b.$$

(b) If $ABC \neq 0$, then setting $a = -\frac{C}{A}$ and $b = -\frac{C}{B}$ (Fig. 3.9), we obtain the *intercept form* of the equation of a line

$$\frac{x}{a} + \frac{y}{b} = 1.$$

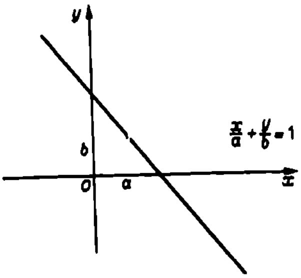


Fig. 3.9

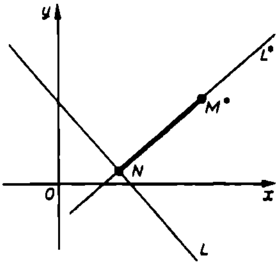


Fig. 3.10

2. The *distance between the point M^* and the line L* is the length of the perpendicular segment M^*N dropped from M^* onto L (Fig. 3.10).

Let $M^*(x^*, y^*)$ be a given point and

$$x \cos\varphi + y \sin\varphi - p = 0$$

be a normal coordinate equation of L. Then the distance between M^* and L is

$$d = d(M^*, L) = |x^* \cos\varphi + y^* \sin\varphi - p|.$$

◄ Let $\mathbf{r}$ be a position vector of an arbitrary point M on L. Then $\mathbf{r} \cdot \mathbf{n}^0 = p$.

We denote the position vector of M^* by $\mathbf{r}^*$. The difference $\mathbf{r}^* \cdot \mathbf{n}^0 - p$ is equal to the projection of the vector $\mathbf{r}^* - \mathbf{r}$ onto the axis L^* determined by $\mathbf{n}^0$, i.e.,

$$\mathbf{r}^* \cdot \mathbf{n}^0 - p = \mathbf{r}^* \cdot \mathbf{n}^0 - \mathbf{r} \cdot \mathbf{n}^0 = (\mathbf{r}^* - \mathbf{r}) \cdot \mathbf{n}^0 = \mathrm{proj}_{L^*}(\mathbf{r}^* - \mathbf{r}),$$

as shown in Fig. 3.11.

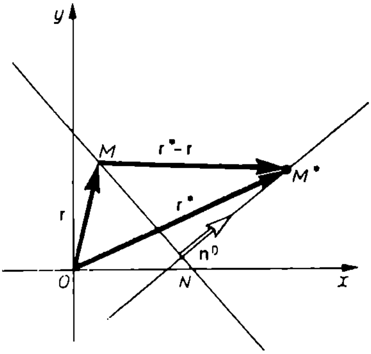


Fig. 3.11

Finding the absolute value of the above identity, we obtain

$$|\mathbf{r}^* \cdot \mathbf{n}^0 - p| = d(M^*, L)$$

or, in terms of coordinates,

$$d(M^*, L) = |x^* \cos\varphi + y^* \sin\varphi - p|. \quad ►$$

If L is specified by the general equation

$$Ax + By + C = 0,$$

the distance between M^* and L is

$$d(M^*, L) = \frac{|Ax^* + By^* + C|}{\sqrt{A^2 + B^2}}.$$

Remark. Similarly, we may define the distance between a point and a plane. If a plane is given by the general equation

$$Ax + By + Cz + D = 0,$$

then the distance between the plane and the point $M^*(x^*, y^*, z^*)$ is

$$d(M^*, \pi) = \frac{|Ax^* + By^* + Cz^* + D|}{\sqrt{A^2 + B^2 + C^2}}.$$

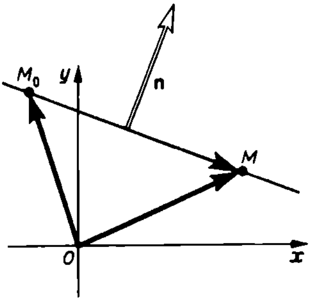


Fig. 3.12

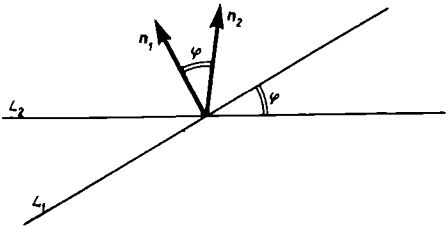


Fig. 3.13

3. The *equation of a line which passes through a given point* $M_0(x_0, y_0)$ *and is perpendicular to a given normal vector* $\mathbf{n} = \{A, B\} \neq 0$ (Fig. 3.12) is

$$A(x - x_0) + B(y - y_0) = 0.$$

4. The *angle between two lines* $A_1x + B_1y + C_1 = 0$ and $A_2x + B_2y + C_2 = 0$ (Fig. 3.13) is given by

$$\cos\varphi = \frac{A_1A_2 + B_1B_2}{\sqrt{A_1^2 + B_1^2}\cdot\sqrt{A_2^2 + B_2^2}}.$$

The two lines are

(a) perpendicular to each other iff

$$A_1A_2 + B_1B_2 = 0,$$

(b) parallel iff

$$\frac{A_1}{A_2} = \frac{B_1}{B_2}.$$

3.3 Straight Line in Three-Dimensional Space

Equation of a line. The line is uniquely determined in space by a position vector $\mathbf{r}_0$ of a point M_0 belonging to the line and a nonzero vector $\mathbf{s}$ parallel to this line. The vector $\mathbf{s}$ is called the *direction vector* of the line.

Let $\mathbf{r} = \overrightarrow{OM}$ be a position vector of an arbitrary point M on the line. From Fig. 3.14 we easily see that

$$\overrightarrow{OM} = \overrightarrow{OM_0} + \overrightarrow{M_0M}. \tag{3.11}$$

The vector $\overrightarrow{M_0M}$ is parallel to the vector $\mathbf{s}$ so that $\overrightarrow{M_0M} = \mathbf{s}t$, where t is a scalar whose value depends on the position of M on the line. Hence, (3.11) may be written as

$$\mathbf{r} = \mathbf{r}_0 + \mathbf{s}t. \tag{3.12}$$

Equation (3.12) is called the *parametric vector equation of a line.*

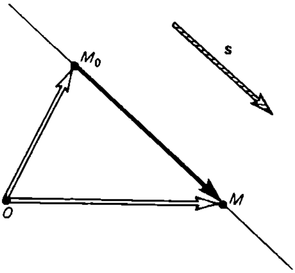


Fig. 3.14

We set up a Cartesian coordinate system with the origin at O. Let us denote the coordinates of M_0, i.e., coordinates of $\mathbf{r}_0$, by x_0, y_0, z_0; the coordinates of an arbitrary point M, i.e., the coordinates of $\mathbf{r}$, by x, y, z; and the coordinates of the direction vector $\mathbf{s}$ by l, m, n. Then the vector equation (3.12) becomes

$$x = x_0 + lt, \quad y = y_0 + mt, \quad z = z_0 + nt. \tag{3.13}$$

Equations (3.13) are called the *parametric equations of a line.* The numbers l, m and n are referred to as *direction numbers* of a line.

Eliminating t from (3.13), we obtain

$$\frac{x - x_0}{l} = t, \quad \frac{y - y_0}{m} = t, \quad \frac{z - z_0}{n} = t.$$

Whence

$$\frac{x - x_0}{l} = \frac{y - y_0}{m} = \frac{z - z_0}{n}, \tag{3.14}$$

where x_0, y_0, z_0 are the coordinates of M_0 belonging to the line and l, m, n are the direction numbers ($l^2 + m^2 + n^2 > 0$).

Relation (3.14) is said to be the *point direction equation* of a line.

Any two equations obtained from (3.14), say

$$\frac{x - x_0}{l} = \frac{y - y_0}{m} \quad \text{and} \quad \frac{y - y_0}{m} = \frac{z - z_0}{n},$$

define a line as a line of intersection of two planes.

Equation of a line passing through two given points. We wish to find equation of a line passing through the points $M_1(x_1, y_1, z_1)$ and $M_2(x_2, y_2, z_2)$. Suppose we are seeking for the point direction equation of a line. So we need to know the coordinates of any point belonging to the line and the direction vector. Let us select $M_1(x_1, y_1, z_1)$ as the point we need and define the direction vector as $\overrightarrow{M_1M_2} = \{x_2 - x_1, y_2 - y_1, z_2 - z_1\}$. Then, the desired equation is

$$\frac{x - x_1}{x_2 - x_1} = \frac{y - y_1}{y_2 - y_1} = \frac{z - z_1}{z_2 - z_1}. \tag{3.15}$$

Remark. If M_1 and M_2 lie in the xy-plane, i.e., if $z_1 = z_2 = 0$, the equation of a line passing through M_1 and M_2 is

$$\frac{x - x_1}{x_2 - x_1} = \frac{y - y_1}{y_2 - y_1}.$$

Example. Find the equation of a line passing through $M_1(1, 0, -1)$ and $M_2(3, 1, 1)$.

◄ Using (3.15), we obtain the desired equation as

$$\frac{x - 1}{2} = \frac{y}{1} = \frac{z + 1}{2}. \quad ►$$

General equation of a line. Reducing the general equation to the point direction equation of a line. A line in space may be determined as a line of intersection of any two distinct nonparallel planes.

Let two distinct nonparallel planes be specified by the equations $A_1x + B_1y + C_1z + D_1 = 0$ and $A_2x + B_2y + C_2z + D_2 = 0$.

Then the system of the two equations

$$\begin{cases} A_1x + B_1y + C_1z + D_1 = 0, \\ A_2x + B_2y + C_2z + D_2 = 0 \end{cases} \tag{3.16}$$

is said to represent the *general equation of a line.*

System (3.16) may be reduced to the point direction equation of a line. This reduction requires a point belonging to the line and a direction vector be known.

We may find the coordinates of a point belonging to the line from (3.16) by setting one unknown equal to an arbitrary value and solving the system of two equations thus obtained to get the other two coordinates of the point we are seeking for. Notice that the direction vector s being of the same direction and lying in the line of intersection of the given planes must be perpendicular to the normal vectors $\mathbf{n}_1 = \{A_1, B_1, C_1\}$ and $\mathbf{n}_2 = \{A_2, B_2, C_2\}$. Conversely, any vector perpendicular to $\mathbf{n}_1$ and $\mathbf{n}_2$ is parallel to both planes, i.e., this vector is parallel to the line of intersection of the planes. The vector product $\mathbf{n}_1 \times \mathbf{n}_2$ is known to be perpendicular both to $\mathbf{n}_1$ and $\mathbf{n}_2$. Hence, we may take $\mathbf{s} = \mathbf{n}_1 \times \mathbf{n}_2$ as the direction vector of the plane.

Example. Reduce

$$\begin{cases} x - y + z - 3 = 0, \\ x + y - 2z + 1 = 0 \end{cases} \tag{3.17}$$

to the point direction equation of a line.

◀ We find a point belonging to the line (3.17). Setting, for example, $z = 0$, we arrive at the system

$$\begin{cases} x - y = 3, \\ x + y = -1. \end{cases}$$

Whence, $x = 1$ and $y = -2$. Thus $M_0(1, -2, 0)$ belongs to the line. ▶

The normal vectors to the planes are $\mathbf{n}_1 = \{1, -1, 1\}$ and $\mathbf{n}_2 = \{1, 1, -2\}$. Then the direction vector of the line is

$$\mathbf{s} = \mathbf{n}_1 \times \mathbf{n}_2 = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & -1 & 1 \\ 1 & 1 & -2 \end{vmatrix} = \mathbf{i} + 3\mathbf{j} + 2\mathbf{k}.$$

Hence, the point direction equation of a line specified by the system (3.17) is

$$\frac{x-1}{1} = \frac{y+2}{3} = \frac{z}{2}. \quad ▶$$

Angle between a line and a plane. Let there be given a line

$$\frac{x - x_0}{l} = \frac{y - y_0}{m} = \frac{z - z_0}{n} \tag{3.18}$$

and a plane

$$Ax + By + Cz + D = 0. \tag{3.19}$$

The angle φ between the line and the plane is the smallest angle between the line and its projection onto the plane (Fig. 3.15).

Let α be an angle between the line (3.18) and the vector normal to the plane (3.19). Then

$$\cos\alpha = \frac{\mathbf{n}\cdot\mathbf{s}}{|\mathbf{n}|\cdot|\mathbf{s}|},$$

where s is the direction vector of the line (3.18) (see Fig. 3.16).

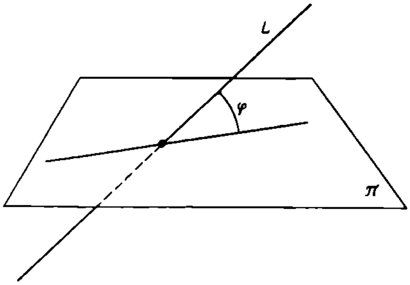


Fig. 3.15

Observe that $|\cos\alpha| = \sin\varphi$. Then

$$\sin\varphi = \frac{|Al + Bm + Cn|}{\sqrt{A^2 + B^2 + C^2}\cdot\sqrt{l^2 + m^2 + n^2}}. \tag{3.20}$$

When the line (3.18) is parallel to the plane (3.19), the direction vector s of the line is perpendicular to the vector **n** normal to the plane so that

$$\mathbf{n}\cdot\mathbf{s} = 0$$

or

$$Al + Bm + Cn = 0. \tag{3.21}$$

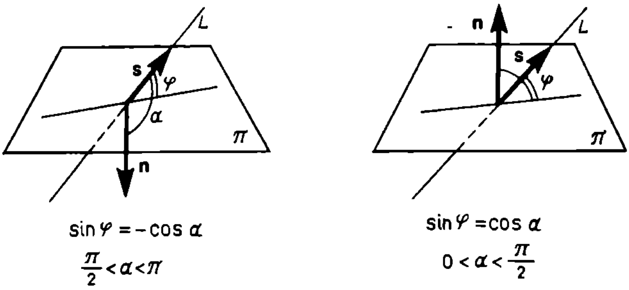

L
s
φ
α
π
n
$\sin\varphi = -\cos\alpha$
$\frac{\pi}{2} < \alpha < \pi$
n
L
s
φ
π
$\sin\varphi = \cos\alpha$
$0 < \alpha < \frac{\pi}{2}$

Fig. 3.16

This is the *condition of parallelism of a line and a plane.*

When the line (3.18) is perpendicular to the plane (3.19), the vectors **s** and **n** are parallel, i.e., $\mathbf{s} \parallel \mathbf{n}$ and

$$\frac{A}{l} = \frac{B}{m} = \frac{C}{n}. \tag{3.22}$$

This is the condition for a line and a plane to be *perpendicular.*

Intersection of a line and a plane. Let there be given a line

$$\frac{x - x_0}{l} = \frac{y - y_0}{m} = \frac{z - z_0}{n} \tag{3.23}$$

and a plane

$$Ax + By + Cz + D = 0. \tag{3.24}$$

The coordinates of the point at which the line (3.23) and the plane (3.24) intersect are to satisfy both (3.23) and (3.24). Thus, we must solve the system containing (3.23) and (3.24) in three unknowns x, y, z to obtain the point at which the line and the plane meet.

We reduce equation (3.23) to the parametric equations

$$x = x_0 + lt, \quad y = y_0 + mt, \quad z = z_0 + nt. \tag{3.25}$$

Substituting (3.25) into (3.24), we obtain

$$Ax_0 + By_0 + Cz_0 + D + t(Al + Bm + Cn) = 0.$$

Whence

$$t = -\frac{Ax_0 + By_0 + Cz_0 + D}{Al + Bm + Cn} \quad (Al + Bm + Cn \neq 0). \tag{3.26}$$

Substituting (3.26) into (3.25), we obtain the coordinates of the desired intersection point.

If $Al + Bm + Cn = 0$ and $Ax_0 + By_0 + Cz_0 + D \neq 0$, the line (3.23) is parallel to the plane (3.24) and the point (x_0, y_0, z_0) does not belong to the plane. Hence, the line (3.23) and the plane (3.24) have no points in common.

When $Al + Bm + Cn = 0$ and $Ax_0 + By_0 + Cz_0 + D = 0$, the former equality implies that the line and the plane are parallel to each other and by virtue of the latter equality the point (x_0, y_0, z_0) belongs to the plane (3.24). Hence, the line (3.23) is contained in the plane (3.24).

Example. Find the point of intersection of the line

$$\frac{x - 7}{3} = \frac{y - 3}{1} = \frac{z + 1}{-2} \tag{3.27}$$

with the plane

$$2x + y + 7z - 3 = 0. \tag{3.28}$$

◄ We reduce (3.27) to the parametric equations

$$x = 7 + 3t, \quad y = 3 + t, \quad z = -1 - 2t. \tag{3.29}$$

Then substituting (3.29) into (3.28), we obtain

$$6t + 14 + t + 3 - 14t - 7 - 3 = 0$$

or

$$-7t + 7 = 0.$$

Whence, $t = 1$.

Now we get from (3.29) the coordinates $x = 10$, $y = 4$, $z = -3$ of the intersection point. ►

Problem. Find the distance between the point $M_1(x_1, y_1, z_1)$ and the line

$$\frac{x - x_0}{l} = \frac{y - y_0}{m} = \frac{z - z_0}{n}.$$

Exercises

1. Write the equation of (a) the plane which is parallel to the xz-plane and passes through the point $(2, -5, 3)$; (b) the plane which contains the z-axis and passes through the point $(-3, 1, -2)$.
2. Given the points $A(1, 3, -2)$ and $B(7, -4, 4)$, write the equation of a plane which passes through B and is perpendicular to the segment AB.
3. Write the equations of (a) the plane which passes through the point $(-2, 7, 3)$ and is parallel to the plane $x - 4y + 5z - 1 = 0$; (b) the plane which passes through the origin of coordinates and is perpendicular to the planes $2x - y + 5z + 3 = 0$ and $x + 3y - z - 7 = 0$.
4. Write the equation of the plane, such that the point $P(3, -6, 2)$ is the foot of the perpendicular dropped from the origin of coordinates onto the plane.
5. Write the equation of the plane passing through the origin of coordinates and the points $A(3, -2, 1)$ and $B(1, 4, 0)$.
6. Write the equation of the line passing through the origin of coordinates and the point (a, b, c).
7. Compute the angle between the lines

$$\frac{x - 1}{3} = \frac{y + 2}{6} = \frac{z - 5}{2} \quad \text{and} \quad \frac{x}{2} = \frac{y - 3}{9} = \frac{z + 1}{6}.$$

(The *angle between two given lines* is that between any two lines which pass through an arbitrary point in space and are parallel to the given lines.)

8. Reduce the equations of the line

$$\begin{cases} 2x - 3y - 3z + 9 = 0, \\ x - 2y + z + 3 = 0 \end{cases}$$

to the point direction equation.

9. Write the equation of the line which passes through the point $(2, -5, 3)$ and is parallel to (a) the z-axis, (b) the line $\dfrac{x-1}{4} = \dfrac{y-2}{-6} = \dfrac{z+3}{9}$,
(c) the line

$$\begin{cases} 2x - y + 3z - 1 = 0, \\ 5x + 4y - z - 7 = 0. \end{cases}$$

10. Write the equation of the line which passes through the point $(1, -1, 0)$ and is perpendicular to the plane $2x - 3y + 5z - 7 = 0$.

11. Find the point of intersection of (a) the line $\dfrac{x-7}{5} = \dfrac{y-4}{1} = \dfrac{z-5}{4}$ with the plane $3x - y + 2z - 5 = 0$; (b) the line $\dfrac{x+1}{2} = \dfrac{y-3}{4} = \dfrac{z}{3}$ with the plane $3x - 3y + 2z - 5 = 0$.

12. Write the equation of the plane which passes through the origin of coordinates and is perpendicular to the line $\dfrac{x+2}{4} = \dfrac{y-3}{5} = \dfrac{z-1}{-2}$.

13. Find the projection of the point $A(4, -3, 1)$ onto the plane $x + 2y - z - 3 = 0$.

14. Write the equation of the plane which passes through the point $(4, -3, 1)$ and is parallel to the lines

$$\frac{x}{6} = \frac{y}{2} = \frac{z}{-3} \quad \text{and} \quad \frac{x+1}{5} = \frac{y-3}{4} = \frac{z-4}{2}.$$

15. Write the equation of the plane which passes through the line $\dfrac{x-3}{2} = \dfrac{y+4}{1} = \dfrac{z-2}{-3}$ and is parallel to the line $\dfrac{x+5}{4} = \dfrac{y-2}{7} = \dfrac{z-1}{2}$.

Answers

1. (a) $y + 5 = 0$; (b) $x + 3y = 0$. **2.** $6x - 7y + 6z - 94 = 0$. **3.** (a) $x - 4y + 5z + 15 = 0$; (b) $2x - y - z = 0$; **4.** $3x - 6y + 2z - 49 = 0$. **5.** $4x - y - 14z = 0$. **6.** $\frac{x}{a} = \frac{y}{b} = \frac{z}{c}$. **7.** $\cos \varphi = \frac{72}{77}$. **8.** $\frac{x}{9} = \frac{y}{5} = \frac{z+3}{-1}$; provided that the point lying on the line is $(0, 0, -3)$. **9.** (a) $\frac{x-2}{0} = \frac{y+5}{0} = \frac{z-3}{1}$ or $\begin{cases} x - 2 = 0, \\ y + 5 = 0 \end{cases}$ (b) $\frac{x-2}{4} = \frac{y+5}{-6} = \frac{z-3}{9}$; (c) $\frac{x-2}{-11} = \frac{y+5}{17} = \frac{z-3}{13}$. **10.** $\frac{x-1}{2} = \frac{y+1}{-3} = \frac{z}{5}$.

11. (a) $(2, 3, 1)$; (b) the line is parallel to the plane. **12.** $4x + 5y - 2z = 0$. **13.** $(5, -1, 0)$. **14.** $16x - 27y + 14z - 159 = 0$. **15.** $23x - 16y + 10z - 153 = 0$.

Chapter 4

Curves and Surfaces of the Second Order

4.1 Changing the Axes of Coordinates in a Plane

Suppose that we are given two Cartesian coordinate systems Oxy and $O'x'y'$ in a plane (Fig. 4.1). We may indicate the location of an arbitrary point M in a plane by giving its coordinates (x, y) on the x- and y-axes or its coordinates (x', y') on the x'- and y'-axes. It is clear that the coordinates (x, y) and (x', y') are somehow related. We shall try to ascertain relationships between (x, y) and (x', y') and, consequently, those between the axes of the two Cartesian coordinate systems in a plane.

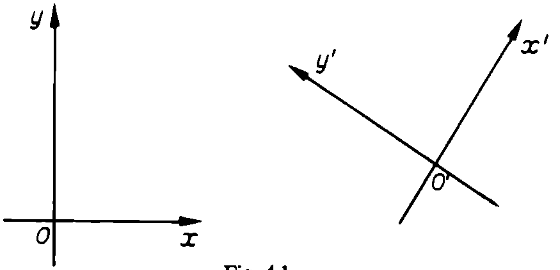


Fig. 4.1

Translation of the axes of coordinates. Suppose that the Cartesian coordinate system $O'x'y'$ is obtained by moving the x- and y-axes parallel to themselves so that the origin of coordinates O goes into the point O' (Fig. 4.2). This means that the unit vectors of the parallel axes are equivalent.

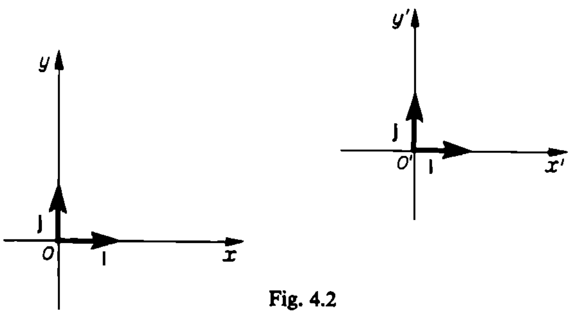


Fig. 4.2

Let **r** and **r**′ be position vectors of a point M relative to the origins O and O' of the given Cartesian coordinate systems Oxy and $O'x'y'$ (Fig. 4.3). Then we have

$$\mathbf{r} = x\mathbf{i} + y\mathbf{j},$$
$$\mathbf{r}' = x'\mathbf{i} + y'\mathbf{j},$$

and

$$\overrightarrow{OO'} = \alpha\mathbf{i} + \beta\mathbf{j},$$

where α and β are the coordinates of the origin O' on the x- and y-axes, respectively.

Since

$$\mathbf{r} = \mathbf{r}' + \overrightarrow{OO'},$$

we obtain

$$x\mathbf{i} + y\mathbf{j} = (x'\mathbf{i} + y'\mathbf{i}) + (\alpha\mathbf{i} + \beta\mathbf{j}).$$

Whence

$$x = x' + \alpha,$$
$$y = y' + \beta.$$

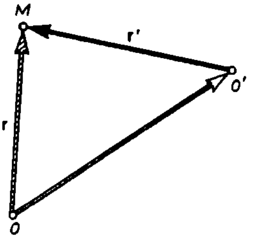


Fig. 4.3

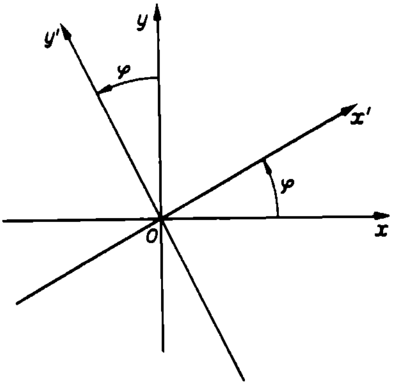


Fig. 4.4

Rotation of the axes of coordinates. Suppose that the x'- and y'-axes are obtained by rotating the x- and y-axes through an angle φ so that the origins of the two Cartesian coordinate systems coincide (Fig. 4.4).

We find coordinates of the unit vectors **i**′ and **j**′ relative to the Cartesian coordinate system Oxy (Fig. 4.5). It is easy to see that the coordinates of the unit vector **i**′ are the cosines of the angles φ and $\frac{\pi}{2} - \varphi$ made by the unit vector **i**′ with the x- and y-axes, respectively. Thus, we may write

$$\mathbf{i}' = \mathbf{i}\cos\varphi + \mathbf{j}\sin\varphi.$$

Similarly, the coordinates of the unit vector $\mathbf{j}'$ are the cosines of the angles $\varphi + \frac{\pi}{2}$ and φ so that

$$\mathbf{j}' = -\mathbf{i}\sin\varphi + \mathbf{j}\cos\varphi.$$

Since for an arbitrary point M its position vectors $\mathbf{r} = x\mathbf{i} + y\mathbf{j}$ and $\mathbf{r}' = x'\mathbf{i}' + y'\mathbf{j}'$ are equivalent, i.e.,

$$x\mathbf{i} + y\mathbf{j} = x'\mathbf{i}' + y'\mathbf{j}',$$

then substituting the expressions for $\mathbf{i}'$ and $\mathbf{j}'$ into the above identity, we obtain

$$\begin{aligned} x\mathbf{i} + y\mathbf{j} &= x'(\mathbf{i}\cos\varphi + \mathbf{j}\sin\varphi) + y'(-\mathbf{i}\sin\varphi + \mathbf{j}\cos\varphi) \\ &= (x'\cos\varphi - y'\sin\varphi)\mathbf{i} + (x'\sin\varphi + y'\cos\varphi)\mathbf{j}. \end{aligned}$$

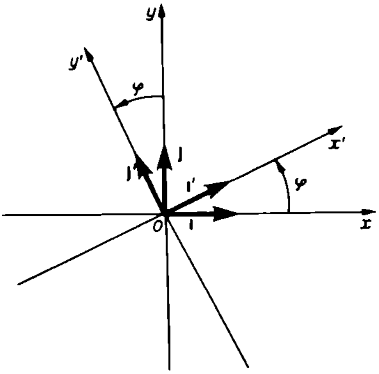


Fig. 4.5

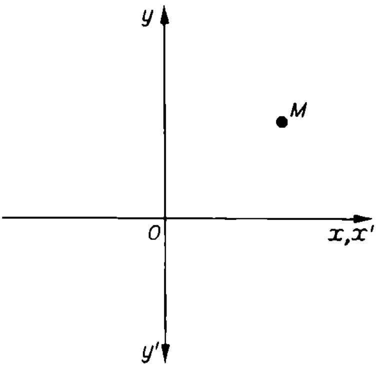


Fig. 4.6

Whence

$$x = x'\cos\varphi - y'\sin\varphi,$$

$$y = x'\sin\varphi + y'\cos\varphi.$$

Reflection of the axes of coordinates. Suppose that the Cartesian coordinate system $O'x'y'$ is obtained from the system Oxy by reversing the y-axis as shown in Fig. 4.6. Then for an arbitrary point M its coordinates (x, y) and (x', y') are related as

$$x' = x \quad \text{and} \quad y' = -y.$$

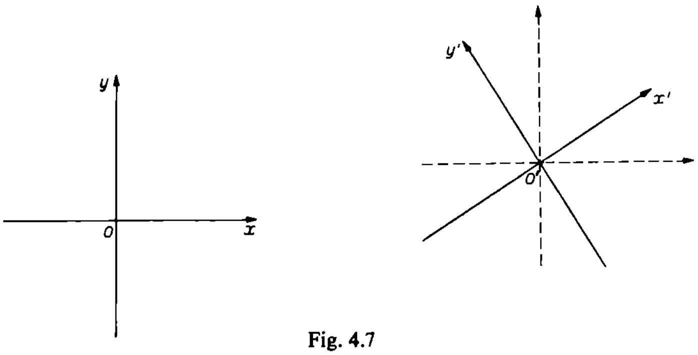


Fig. 4.7

Any change from one Cartesian coordinate system to another one with the same unit distance may be made using in succession the translation, rotation and reflection of the axes (Fig. 4.7).

4.2 Curves of the Second Order

Suppose that we are given a Cartesian coordinate system in a plane. A locus of points of a plane whose coordinates satisfy the equation

$$F(x, y) = 0, \qquad (*)$$

where $F(x, y)$ is a function of two variables, is called a *plane curve.* The equation $(*)$ is said to be the *equation of a plane curve.*

For example, the equation $x + y = 0$ is the equation of a line which bisects the second and the fourth quadrants of the coordinate plane (Fig. 4.8), and the equation $x^2 + y^2 - 1 = 0$ is the equation of a circle with unit radius and centre at the origin of coordinates (Fig. 4.9).

We consider the quadratic polynomial of two variables x and y

$$F(x, y) = Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F$$

$$(A^2 + B^2 + C^2 > 0).$$

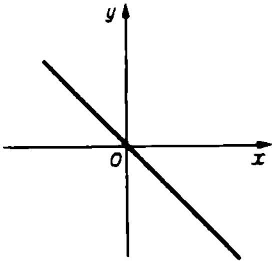


Fig. 4.8

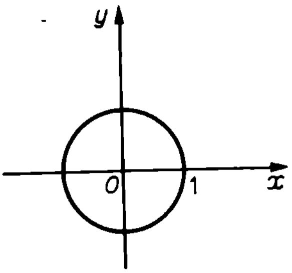


Fig. 4.9

The equation

$$F(x, y) = 0$$

is said to be the *equation of a curve of the second order.*

While curves of the first order are straight lines (and only straight lines), the quadratic polynomial equation provides a variety of curves of the second order. So it is helpful to precede our study of the general equation of a curve of the second order with an investigation of some important specific plane curves of the second order.

4.3 The Ellipse

The *ellipse* is a plane curve defined by the Cartesian equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \tag{4.1}$$

where $a \geqslant b > 0$.

Equation (4.1) is called the ***standard Cartesian equation of the ellipse*** and the associated Cartesian coordinate system is called the ***standard Cartesian coordinate system.***

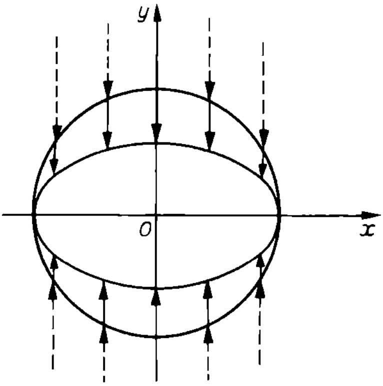


Fig. 4.10

The circle

$$x^2 + y^2 = a^2 \tag{4.2}$$

is an ellipse with $a = b$. This enables us to regard the ellipse (4.1) as a plane figure obtained by uniformly compressing the circle (4.2) with coefficient b/a towards the x-axis (Fig. 4.10). In other words, equation (4.1) of an ellipse is obtained by substituting $(a/b)y$ for y in the equation $x^2 + y^2 = a^2$. (The uniform compression of a circle towards the x-axis with coefficient $k > 0$ is a transformation which sends an arbitrary point $M(x, y)$ on the circle to the point $M'(x, y/k)$.)

Properties of the ellipse. (1) The ellipse (4.1) is contained in the rectangle

$$P = \{(x, y) : |x| \leqslant a, \quad |y| \leqslant b\}.$$

◀ This property is easy to verify since

$$\frac{x^2}{a^2} \leqslant 1 \quad \text{and} \quad \frac{y^2}{b^2} \leqslant 1$$

for any point $M(x, y)$ lying on the ellipse. ▶

The points $(\pm a, 0)$ and $(0, \pm b)$ are called the *vertices of the ellipse* (Fig. 4.11).

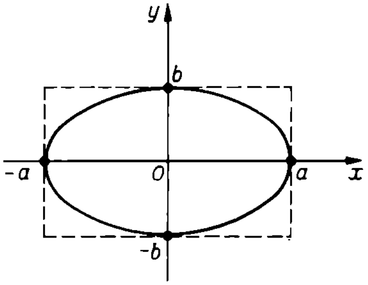


Fig. 4.11

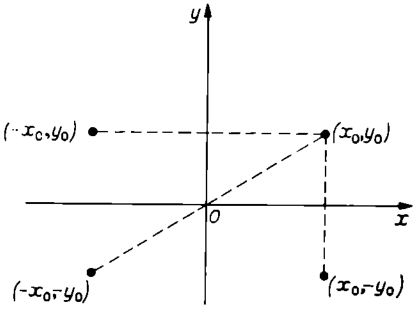


Fig. 4.12

(2) For the ellipse (4.1) the x- and y-axes of the standard Cartesian coordinate system are the axes of symmetry and the origin O of coordinates is the centre of symmetry. This means that when $M_0(x_0, y_0)$ belongs to the ellipse, so do the points $(-x_0, y_0)$, $(-x_0, -y_0)$ and $(x_0, -y_0)$ (Fig. 4.12).

(3) If $a > b$, that is, unless the ellipse is a circle, the x- and y-axes of the standard Cartesian coordinate system are the unique axes of symmetry of the ellipse (4.1).

Set $c = \sqrt{a^2 - b^2}$. It is easy to see that $c < a$. The points $(-c, 0)$ and $(c, 0)$ are called the left-hand and right-hand *foci* of the ellipse. The distance between the foci is equal to $2c$.

(4) The ellipse is a locus of points of a plane such that the sum of the distances of these points from two given points (from the foci of the ellipse) is constant (is equal to a given number) (Fig. 4.13).

◄ Let $M(x, y)$ be a point lying on the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

The distance from $M(x, y)$ to the left-hand and to the right-hand foci of the ellipse are

$$\varrho_l = \sqrt{(x + c)^2 + y^2} \quad \text{and} \quad \varrho_r = \sqrt{(x - c)^2 + y^2}.$$

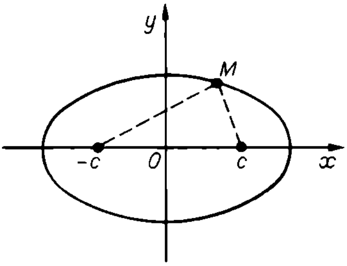


Fig. 4.13

By the substitution

$$y^2 = b^2\left(1 - \frac{x^2}{a^2}\right)$$

we easily obtain

$$\varrho_l = \sqrt{\left(1 - \frac{b^2}{a^2}\right)x^2 + 2xc + c^2 + b^2}$$

$$= \sqrt{\frac{c^2}{a^2}x^2 + 2cx + a^2} = \left|\frac{c}{a}x + a\right| = a + \frac{c}{a}x,$$

because $|x| \leqslant a$ and $\frac{c}{a} < 1$.

Similarly, we obtain

$$\varrho_r = a - \frac{c}{a}x.$$

Adding ϱ_l and ϱ_r, we have the desired result

$$\varrho_l + \varrho_r = 2a.$$

Notice that we have proved in Chapter 1 (see Section 2, Exercise 2) that any point possessing this property lies on the ellipse. ►

The number $e = \frac{c}{a}$ is called the *eccentricity* of the ellipse (4.1). It is easy to see that $0 < e < 1$, and $e = 0$ in the case of a circle.

The lines

$$x + \frac{a}{e} = 0 \quad \text{and} \quad x - \frac{a}{e} = 0$$

are called the *directrices* of the ellipse (4.1) (Fig. 4.14). Any ellipse has two directrices.

(5) The ellipse is a locus of points of a plane such that for any point of the locus the ratio of its distance from a given point (from the focus of the ellipse) to its distance from a given line (from the directrix lying on the same side of the y-axis as the focus) is constant (is equal to the eccentricity of the ellipse).

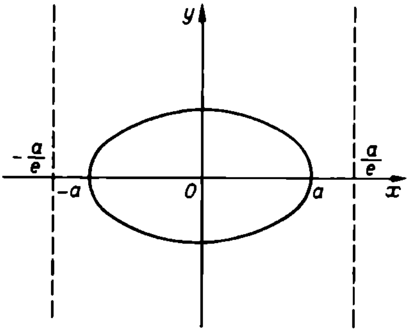


Fig. 4.14

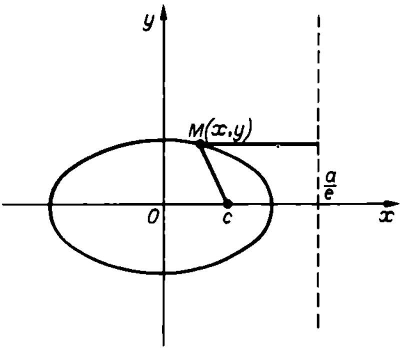


Fig. 4.15

◄ Let $M(x, y)$ be a point on the ellipse (4.1) (Fig. 4.15). The distances from $M(x, y)$ to the right-hand focus and to the right-hand directrix are

$$\varrho_r = a - ex \quad \text{and} \quad d_r = \frac{a}{e} - x.$$

Whence we obtain the desired result

$$\frac{\varrho_r}{d_r} = e.$$

Similarly, we obtain

$$\frac{\varrho_l}{d_l} = \frac{a + ex}{\frac{a}{e} + x} = e.$$

Now we consider the point $(c, 0)$ and the line $x = \frac{a}{e}$ $(c = ae)$. The distance from the point $M(x, y)$ to the given point $(c, 0)$ and to the given line $y = \frac{a}{e}$ are

$$\sqrt{(x-c)^2 + y^2} \quad \text{and} \quad \left|\frac{a}{e} - x\right|,$$

respectively.

We require that

$$\frac{\sqrt{(x-c)^2 + y^2}}{\left|\frac{a}{e} - x\right|} = e.$$

Then

$$\sqrt{(x-c)^2 + y^2} = |a - ex|.$$

Squaring the above identity and setting $b^2 = a^2 - c^2$, we easily obtain

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

Thus the point $M(x, y)$ belongs to the ellipse (4.1). ►

4.4 The Hyperbola

The *hyperbola* is a plane curve given by the Cartesian equation

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1, \tag{4.3}$$

where $a > 0$ and $b > 0$.

Equation (4.3) is called the *standard Cartesian equation of the hyperbola* and the associated Cartesian coordinate system is called the *standard Cartesian coordinate system.*

Properties of the hyperbola. (1) The hyperbola (4.3) lies outside the strip $|x| < a$ (Fig. 4.16).

◄ It is easy to see that for any point lying on the hyperbola there holds

$$\frac{x^2}{a^2} = 1 + \frac{y^2}{b^2} \geqslant 1.$$

Therefore $|x| \geqslant a$ for any point lying on the hyperbola. ►

The points $(\pm a, 0)$ are the *vertices of the hyperbola.*

(2) The hyperbola (4.3) lies in the interior of the vertical angles formed by the lines $y = \pm\frac{b}{a}x$, the x-axis being the bisector of these angles (Fig. 4.17).

◄ From the inequality

$$\frac{x^2}{a^2} > \frac{y^2}{b^2}$$

it follows that for any point $M(x, y)$ lying on the hyperbola

$$|y| < \frac{b}{a}|x|$$

holds. ►

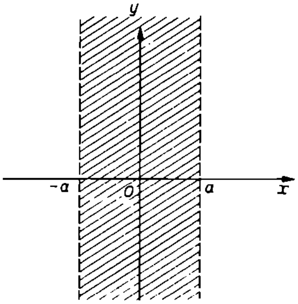


Fig. 4.16

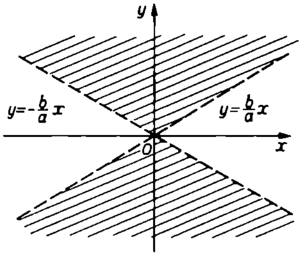


Fig. 4.17

Thus the hyperbola contains two parts called the *branches of the hyperbola.*

The lines

$$\frac{x}{a} + \frac{y}{b} = 0 \quad \text{and} \quad \frac{x}{a} - \frac{y}{b} = 0$$

are called the *asymptotes of the hyperbola.*

(3) The hyperbola contains points infinitely distant from the origin O of coordinates.

◄ For example, let $M(x, y)$ be a point on the hyperbola and let $|y| = n$, where n is an arbitrary positive number (Fig. 4.18).

Then

$$|x| = a\sqrt{1 + \frac{y^2}{b^2}} > \frac{a}{b}|y| = \frac{a}{b}n$$

and

$$\sqrt{x^2 + y^2} \geqslant \frac{|x| + |y|}{2} > n\frac{a + b}{2b}. \quad ►$$

Let us establish relations between points lying on the hyperbola and points belonging to the asymptotes of the hyperbola.

We choose a point M lying on the hyperbola (4.3) in the first quadrant of the standard Cartesian coordinate system and a point N belonging to the asymptote $\frac{x}{a} - \frac{y}{b} = 0$ and having the same abscissa as M (Fig. 4.19). These points are

$$M\left(x,\ b\sqrt{\frac{x^2}{a^2} - 1}\right) \quad \text{and} \quad N\left(x,\ \frac{b}{a}x\right).$$

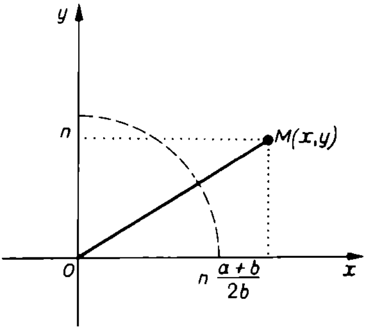


Fig. 4.18

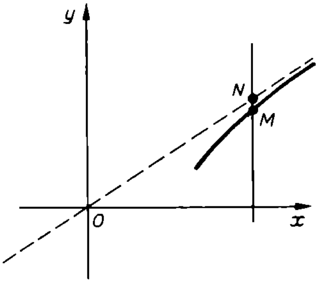


Fig. 4.19

The distance between the points M and N is

$$d(M,\ N) = \frac{b}{a}\left(x - \sqrt{x^2 - a^2}\right).$$

Multiplying and dividing $d(M,\ N)$ by $x + \sqrt{x^2 + a^2}$, we obtain

$$d(M,\ N) = \frac{ab}{x + \sqrt{x^2 - a^2}}.$$

Passing to the limit we conclude that $d(M,\ N)$ tends to zero as x tends to infinity.

Thus we have proved the following property of the hyperbola:

(4) If a point recedes along the asymptote from the origin of coordinates into infinity, i.e., if $x \to +\infty$, then there exists a point lying on the hyperbola and having the same abscissa as the point on the asymptote so that the distance between these two points tends to zero.

The converse is also true, namely:

(5) If a point $M(x,\ y)$ recedes along the hyperbola from the origin of coordinates into infinity, i.e., if $x^2 + y^2 \to \infty$, then the distance from

$M(x, y)$ either to the asymptote

$$\frac{x}{a} + \frac{y}{b} = 0$$

or to the asymptote

$$\frac{x}{a} - \frac{y}{b} = 0$$

tends to zero.

(6) The axes of the standard Cartesian coordinate system are the axes of symmetry for the hyperbola (4.3) and the origin of coordinates is its centre of symmetry (Fig. 4.20).

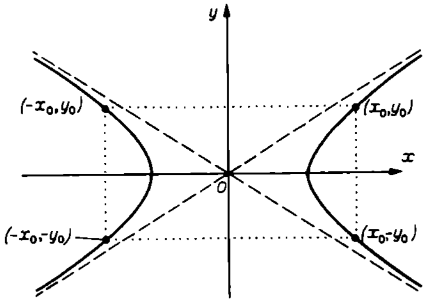


Fig. 4.20

The axes of the standard Cartesian coordinate system are the unique axes of symmetry of the hyperbola.

Set $c = \sqrt{a^2 + b^2}$. It is easy to see that $c > 0$. The points $(-c, 0)$ and $(c, 0)$ are called the *foci of the hyperbola.* The distance between the foci is $2c$.

(7) The hyperbola is a locus of points of a plane such that for any point the absolute value of the difference of its distances to two given points (to the foci of the hyperbola) is constant (is equal to a given number). ◄ The proof is similar to that of Property 4 for the ellipse. For example, we shall show that each point of the hyperbola possesses this property. Let $M(x, y)$ be a point lying on the hyperbola (4.3) (Fig. 4.21). Then the distances from $M(x, y)$ to the foci of the hyperbola are

$$\varrho_l = \sqrt{(x + c)^2 + y^2} = \left| a + \frac{c}{a} x \right|$$

and

$$\varrho_r = \sqrt{(x - c)^2 + y^2} = \left| a - \frac{c}{a} x \right|.$$

Since $\dfrac{c}{a} > 1$,

$$\varrho_l = \begin{cases} a + \dfrac{c}{a}x & \text{if} \quad x \geqslant a, \\ -a - \dfrac{c}{a}x & \text{if} \quad x \leqslant -a \end{cases}$$

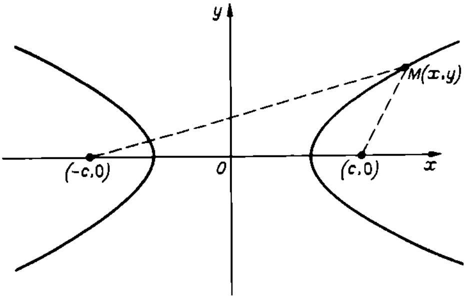


Fig. 4.21

and

$$\varrho_r = \begin{cases} -a + \dfrac{c}{a}x & \text{if} \quad x \geqslant a, \\ a - \dfrac{c}{a}x & \text{if} \quad x \leqslant -a. \end{cases}$$

Whence

$$\varrho_l - \varrho_r = \begin{cases} 2a & \text{if} \quad x \geqslant a, \\ -2a, & \text{if} \quad x \leqslant -a. \end{cases}$$

Therefore we may write

$$|\varrho_l - \varrho_r| = 2a. \quad ►$$

The number $e = \dfrac{c}{a}$ is called the *eccentricity of the hyperbola* (4.3). It is easy to see that $e > 1$.

The lines

$$x + \frac{a}{e} = 0 \quad \text{and} \quad x - \frac{a}{e} = 0$$

are called the *directrices of the hyperbola* (4.3).

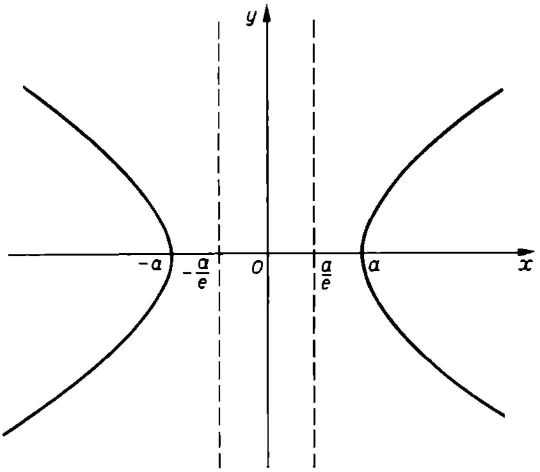


Fig. 4.22

Any hyperbola has two directrices (Fig. 4.22).

The following property is similar to Property 5 of the ellipse:

(8) The hyperbola is a locus of points of a plane such that for any point the ratio of its distance from a given point (from the focus of the hyperbola) to its distance from a given line (from the directrix lying on the same side of the y-axis as the focus) is constant (is equal to the eccentricity of the hyperbola) (Fig. 4.23).

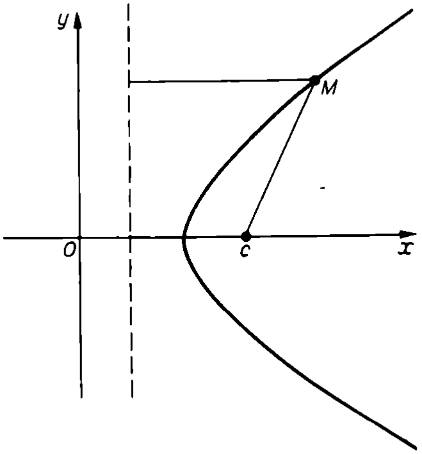


Fig. 4.23

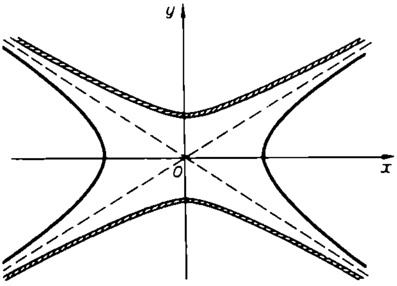


Fig. 4.24

The hyperbola

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = -1 \tag{4.4}$$

is called the *conjugate hyperbola* of the hyperbola (4.3). Locations of the hyperbolas (4.3) and (4.4) are shown in Fig. 4.24.

4.5 The Parabola

The *parabola* is a plane curve defined by the Cartesian equation

$$y^2 = 2px, \tag{4.5}$$

where $p > 0$.

Equation (4.5) is called the *standard Cartesian equation of the parabola* and the associated Cartesian coordinate system is called the *standard Cartesian coordinate system* (Fig. 4.25).

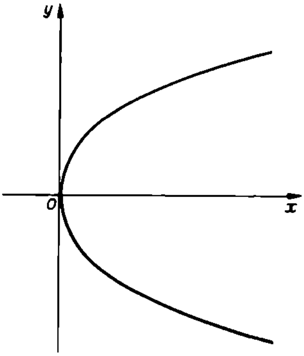


Fig. 4.25

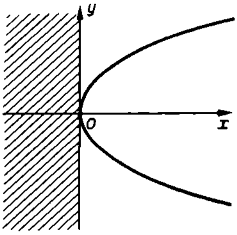


Fig. 4.26

Properties of the parabola. (1) The parabola lies in a semiplane to the right from the y-axis, i.e., $x \geqslant 0$ for any point lying on the parabola (Fig. 4.26). The origin of coordinates belongs to the parabola and is called the *vertex of the parabola* (4.5).

(2) The parabola contains points infinitely distant from the origin of coordinates.

(3) The axis of abscissas of the standard Cartesian coordinate system is the unique axis of symmetry of the parabola (4.5) (Fig. 4.27).

The axis of symmetry is called the *axis of the parabola.*

The number p is sometimes called the *focal parameter* of the parabola. The point $\left(\frac{p}{2}, 0\right)$ is the *focus*, and the line $x = -\frac{p}{2}$ is the *directrix* of the parabola (4.5).

(4) The parabola is a locus of points of a plane such that for any point its distance from a given point (from the focus of the parabola) is equal to its distance from a given line (the directrix of the parabola) (Fig. 4.28).

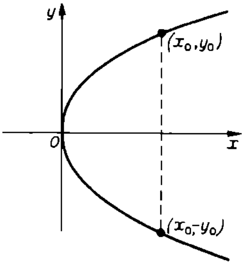


Fig. 4.27

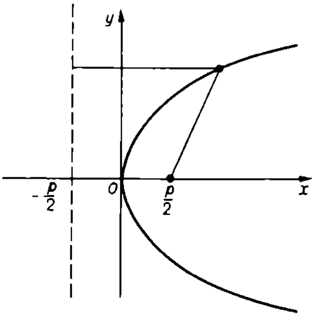


Fig. 4.28

◄ Let $M(x, y)$ be a point lying on the parabola (4.5). The distances from $M(x, y)$ to the focus $\left(\frac{p}{2}, 0\right)$ and from $M(x, y)$ to the directrix $x = -\frac{p}{2}$ are

$$\sqrt{\left(x - \frac{p}{2}\right)^2 + y^2} \quad \text{and} \quad \left|x + \frac{p}{2}\right|,$$

respectively.

Substituting $2px$ for y^2, we easily obtain

$$\sqrt{\left(x - \frac{p}{2}\right)^2 + y^2} = \sqrt{\left(x - \frac{p}{2}\right)^2 + 2px} = \sqrt{\left(x + \frac{p}{2}\right)^2} = \left|x + \frac{p}{2}\right|.$$

The converse is also true Let $M(x, y)$ be a point equidistant from the point $\left(\frac{p}{2}, 0\right)$ and from the line $x = -\frac{p}{2}$, i.e.,

$$\sqrt{\left(x - \frac{p}{2}\right)^2 + y^2} = \left|x + \frac{p}{2}\right|.$$

Squaring the above identity, we easily obtain

$$y^2 = 2px. \blacktriangleright$$

4.6 Optical Properties of Curves of the Second Order

Tangents to ellipses and hyperbolas. Let $y = f(x)$ be an equation of a plane curve. Then the equation of a tangent to the plane curve at a point (x_0, y_0), where $y_0 = f(x_0)$, may be written in the form

$$y - y_0 = f'(x_0)(x - x_0). \tag{4.6}$$

Let $M_0(x_0, y_0)$ be a point lying on the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

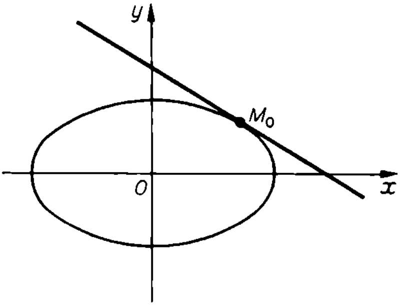


Fig. 4.29

For definiteness we assume that the point M_0 lies in the first quadrant of the coordinate plane, i.e., $x_0 > 0$ and $y_0 > 0$ (Fig. 4.29).

The part of the ellipse in the first quadrant is given by the equation

$$y = b\sqrt{1 - \frac{x^2}{b^2}}.$$

Using (4.6), we obtain the equation of a tangent to the ellipse at the point $M_0(x_0, y_0)$

$$y - y_0 = -\frac{x_0 b}{a^2\sqrt{1 - \frac{x_0^2}{a^2}}}(x - x_0).$$

Since the point $M_0(x_0, y_0)$ belongs to the ellipse,

$$y_0 = b\sqrt{1 - \frac{x_0^2}{a^2}}.$$

Whence we have

$$y - y_0 = -\frac{x_0 b^2}{a^2 y_0}(x - x_0).$$

The above relation may be presented in the form

$$\frac{xx_0}{a^2} + \frac{yy_0}{b^2} - \left(\frac{x_0^2}{a^2} + \frac{y_0^2}{b^2}\right) = 0.$$

Since

$$\frac{x_0^2}{a^2} + \frac{y_0^2}{b^2} = 1,$$

we arrive at the equation

$$\frac{xx_0}{a^2} + \frac{yy_0}{b^2} = 1.$$

This is the equation of the tangent to the ellipse at the point (x_0, y_0) which is independent of the location of M_0 in the coordinate plane. In other words, this equation is satisfied by the Cartesian coordinates of each point lying on the ellipse.

Similarly, we may easily derive the equation of a tangent to the hyperbola in the form

$$\frac{xx_0}{a^2} - \frac{yy_0}{b^2} = 1,$$

the point (x_0, y_0) lying on the hyperbola.

Tangents to parabolas. Let $x = g(y)$ be an equation of a plane curve. Then the equation of a tangent to this plane curve at a point (x_0, y_0), where $x_0 = g(y_0)$, may be written in the form

$$x - x_0 = g'(y_0)(y - y_0). \tag{4.7}$$

Let $M_0(x_0, y_0)$ be a point lying on a parabola. Using (4.7), we obtain the equation of a tangent to the parabola at $M_0(x_0, y_0)$ in the form

$$x - x_0 = \frac{y_0}{p}(y - y_0)$$

or

$$yy_0 - y_0^2 + px_0 - px = 0.$$

Whence, observing that $y_0^2 = 2px_0$, we arrive at the equation of a tangent of the form

$$yy_0 = p(x + x_0).$$

Remark. Comparing standard Cartesian equations of the ellipse, hyperbola and parabola with the equations of respective tangents we see that the latter can easily be obtained from the former. Indeed, substituting yy_0 for y^2 and xx_0 for x^2 in the equations of the ellipse and hyperbola, and $x + x_0$ for $2x$ in the equation of the parabola, we immediately arrive at the equations of the respective tangents. Notice that the point with coordinates (x_0, y_0) lies on a plane curve.

Optical property of the ellipse. Let $M_0(x_0, y_0)$ be a point lying on the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

Recall that the distance from M_0 to the foci F_l and F_r of the ellipse are

$$\varrho_l = |ex_0 + a| \quad \text{and} \quad \varrho_r = |ex_0 - a|,$$

respectively.

We draw through M_0 a tangent line given by the equation

$$\frac{xx_0}{a^2} + \frac{yy_0}{b^2} = 1.$$

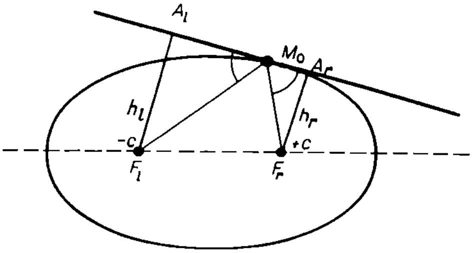


Fig. 4.30

It is easy to find the distances from the tangent to the foci $F_l(-c, 0)$ and $F_r(c, 0)$ (Fig. 4.30). We have

$$h_l = \mu \left| \frac{cx_0}{a^2} + 1 \right| \quad \text{and} \quad h_r = \mu \left| \frac{cx_0}{a^2} - 1 \right|,$$

where $\mu = 1\Big/\sqrt{\dfrac{x_0^2}{a^4} + \dfrac{x_0^2}{b^4}}$ is a normalizing factor.

We can easily verify that

$$\frac{h_l}{\varrho_l} = \frac{h_r}{\varrho_r}.$$

Indeed,

$$\frac{h_l}{\varrho_l} = \frac{\mu \left| \dfrac{cx_0}{a^2} + 1 \right|}{|ex_0 + a|} = \frac{\mu}{a} \frac{|ex_0 + a|}{|ex_0 + a|} = \frac{\mu}{a}$$

and

$$\frac{h_r}{\varrho_r} = \frac{\mu \left| \dfrac{cx_0}{a^2} - 1 \right|}{|ex_0 - a|} = \frac{\mu}{a}.$$

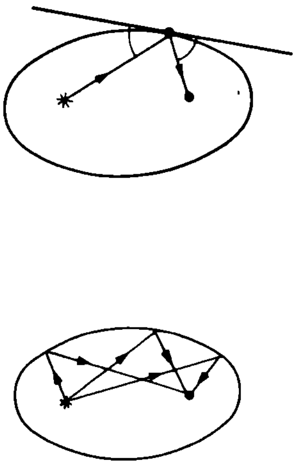


Fig. 4.31

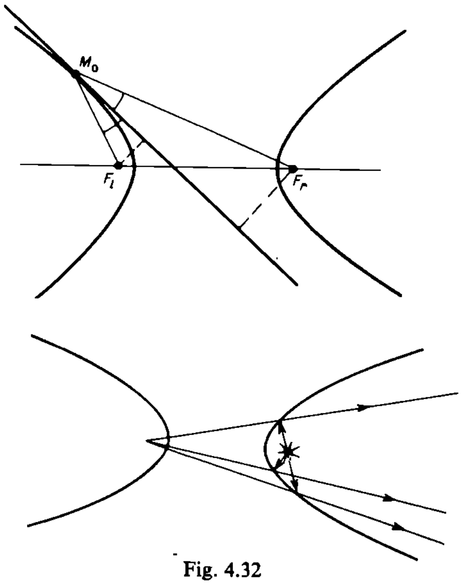


Fig. 4.32

We see from Fig. 4.30 that the above ratios are equal to the sines of the angles made by the segments F_lM_0 and F_rM_0 with the tangent. Since the sines are equal so are the angles. Therefore a tangent at any point lying on an ellipse makes equal angles with the line segments joining this point with the foci of the ellipse. This property is sometimes called the *optical property of the ellipse* because the rays from a source of light placed at one of the foci are reflected from the "mirror" surface of the ellipse and are focused into the other focus (Fig. 4.31).

Optical property of the hyperbola. Similarly, we may establish the following optical property of the hyperbola: the rays from a source of light placed at one of the foci seem to emanate from the other focus after being reflected from a "mirror" surface of the hyperbola (Fig. 4.32).

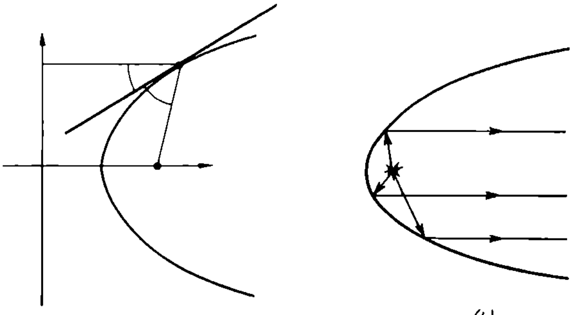

(a)
(b)

Fig. 4.33

Optical property of the parabola. If a source of light is placed at the focus of a parabola the rays from this source are reflected by the "parabolic mirror" in directions parallel to the axis of the parabola (Fig. 4.33).

4.7 Classification of Curves of the Second Order

We begin our classification with the discussion of plane curves given by *quadratic polynomials.*

Theorem 4.1. *Let us set up a Cartesian coordinate system Oxy in a plane and let*

$$f(x, y) = ax^2 + 2bxy + cy^2 + 2dx + 2ey + g$$
$$(a^2 + b^2 + c^2 > 0) \tag{4.8}$$

be a quadratic polynomial of two variables x and y.

Then there exists a Cartesian coordinate system O'XY such that substituting X for x and Y for y reduces the polynomial f(x, y) to the polynomial F(X, Y) which may be of the following three kinds:

(a) $AX^2 + BY^2 + C, \quad A \cdot B \neq 0;$

(b) $BY^2 + 2DX, \quad B \cdot D \neq 0;$

(c) $BY^2 + E, \quad B \neq 0.$

◄ *Step 1.* By a suitable rotation of the x- and y-axes we may eliminate the term $2bxy$ from $f(x, y)$.

Let $b \neq 0$.

Consider a Cartesian coordinate system $Ox'y'$ obtained from the original system Oxy by rotating the x- and y-axes through an angle φ (Fig. 4.34). The new and original axes of coordinates are related as

$$\begin{aligned} x &= x' \cos\varphi - y' \sin\varphi, \\ y &= x' \sin\varphi + y' \cos\varphi. \end{aligned} \tag{4.9}$$

Substituting (4.9) into (4.8), we reduce the term $2bxy$ to the term $2b'x'y'$, where

$$\begin{aligned} 2b' &= 2(-a \sin\varphi \cos\varphi + b(\cos^2\varphi - \sin^2\varphi) + c \sin\varphi \cos\varphi) \\ &= (c - a)\sin 2\varphi + 2b \cos 2\varphi. \end{aligned}$$

To eliminate $2b'x'y'$ from the polynomial $f(x', y')$ it suffices to set $2b' = 0$. Whence we have

$$\cot 2\varphi = \frac{a - c}{2b}.$$

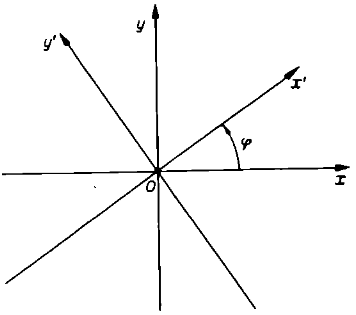


Fig. 4.34

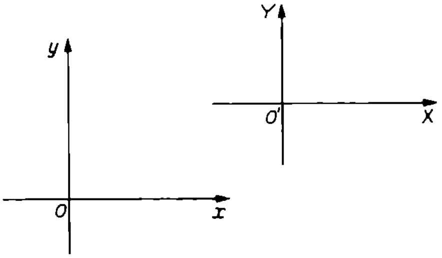


Fig. 4.35

Observe that a, b and c are known. From the above identity we may find the value φ of the angle through which we should rotate the original axes of coordinates to eliminate the term $2b'x'y'$ from the polynomial $f(x', y')$. In other words, we may always choose a Cartesian coordinate system such that the original polynomial is reduced to the form

$$f(x, y) = ax^2 + cy^2 + 2dx + 2ey + g,$$

where $a^2 + c^2 > 0$, and in what follows we shall consider the above polynomial. For definiteness, we also set $c \neq 0$, since substituting y for x, we may always reduce the polynomial to a form with $c \neq 0$.

Step 2. By a suitable translation of the axes of coordinates we may further simplify the polynomial $f(x, y)$.

Consider a Cartesian coordinate system $O'XY$ obtained from the system Oxy by translating the x- and y-axes (Fig. 4.35) in which case

$$X = x + \alpha,$$
$$Y = y + \beta, \tag{4.10}$$

where $-\alpha$ and $-\beta$ are the coordinates of the origin O' of the system $O'XY$ in the original system Oxy.

Depending on the relations between the coefficients a, b, c, d and e, the polynomial $f(x, y)$ reduces to one of the following three kinds

(1) $a \neq 0$, $c \neq 0$. Then setting $\alpha = \frac{d}{a}$ and $\beta = \frac{e}{c}$, we obtain

$$F(X, Y) = AX^2 + BY^2 + C,$$

where $A = a$, $B = b$, and $C = g - \frac{d^2}{a} - \frac{e^2}{c}$.

(2) $a = 0$, $d \neq 0$. Then setting

$$\alpha = \frac{1}{2d}\left(g - \frac{e^2}{c}\right) \quad \text{and} \quad \beta = \frac{e}{c}$$

we obtain

$$F(X, Y) = BY^2 + 2DX,$$

where $B = c$ and $D = d$.

(3) $a = d = 0$. Then setting

$$\alpha = 0 \quad \text{and} \quad \beta = \frac{e}{c}$$

we obtain

$$F(X, Y) = BY^2 + E,$$

where $B = c$ and $E = g - \frac{e^2}{c}$. ►

Standard equations of curves of the second order. We consider plane curves of the second order defined by the polynomial equation

$$F(X, Y) = 0,$$

where $F(X, Y)$ is a quadratic polynomial discussed in the preceding theorem.

Depending on the kind of $F(X, Y)$ the polynomial equation determines equations of plane curves of different shapes. Let us consider them separately.

(1) $F(X, Y) = AX^2 + BY^2 + C = 0,\ A \cdot B \neq 0.$

Assume that $A \cdot B > 0$. Multiplying the polynomial equation by (-1) and substituting Y for X and X for Y, if necessary, we may reduce the equation to a form with $B \geqslant A > 0$. Then

(a) If $C < 0$, we obtain an *equation of the ellipse*

$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} = 1,$$

where $a^2 = -\frac{C}{A}$, $b^2 = -\frac{C}{B}$ (a $\geqslant b > 0$).

(b) If $C > 0$, we obtain an *equation of the imaginary ellipse**)

$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} = -1,$$

where $a^2 = \frac{C}{A}$ and $b^2 = \frac{C}{B}$ $(a \geqslant b > 0)$.

Notice that in the XY-plane there exists no point whose coordinates satisfy this equation.

(c) If $C = 0$, we obtain the equation

$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} = 0,$$

where $a^2 = \frac{1}{A}$ and $b^2 = \frac{1}{B}$ (a $\geqslant b > 0$).

The origin of coordinates is the only point whose coordinates satisfy this equation. We may regard the origin of coordinates as a point of intersection of *two imaginary lines***).

Assume that $A \cdot B < 0$. Multiplying the polynomial equation by (-1) and substituting Y for X and, if necessary, X for Y we may always reduce the equation to a form with $A > 0$, $B < 0$ and $C \leqslant 0$.

Then

(a) If $C < 0$, we obtain an equation of the *hyperbola*

$$\frac{X^2}{a^2} - \frac{Y^2}{b^2} = 1,$$

where $a^2 = -\frac{C}{A}$ and $b^2 = \frac{C}{B}$ $(a > 0,\ b > 0)$.

---

*) This curve is called the imaginary ellipse since its equation resembles the equation of the real ellipse.

**) We speak of imaginary lines since the respective equation resembles the equation defining two intersecting lines in a plane.

(b) If $C = 0$, we obtain the equation

$$\frac{X^2}{a^2} - \frac{Y^2}{b^2} = 0,$$

where $a^2 = \frac{1}{A}$ and $b^2 = -\frac{1}{B}$.

This equation defines *two intersecting lines*

$$\frac{X}{a} - \frac{Y}{b} = 0 \quad \text{and} \quad \frac{X}{a} + \frac{Y}{b} = 0$$

in a plane.

(2) $F(X, Y) = BY^2 + 2DX = 0$, $B \cdot D \neq 0$.

Assume that $B \cdot D < 0$. Substituting if necessary $-X$ for X we always reduce the polynomial equation to a form with $B \cdot D < 0$. Then we obtain an equation of the *parabola*

$$Y^2 = 2pX,$$

where $p = -\frac{D}{B}$ $(p > 0)$.

(3) $F(X, Y) = BY^2 + E = 0$, $B \neq 0$.

Assume that $B > 0$. Then

(a) If $E < 0$, we obtain the equation

$$Y^2 - c^2 = 0,$$

where $c^2 = -\frac{E}{B}$ $(c > 0)$.

This equation defines *two parallel lines in a plane.*

(b) If $E > 0$, we obtain the equation

$$Y^2 + c^2 = 0,$$

where $c^2 = \frac{E}{B}$ $(c > 0)$.

There is no plane which contains a point whose coordinates satisfy this equation, called the *equation of two imaginary parallel lines* because it resembles the equation defining two parallel lines.

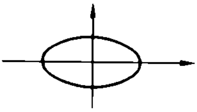
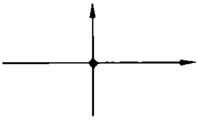
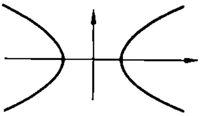
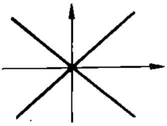
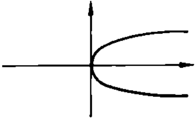
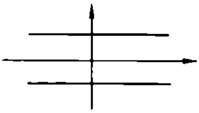
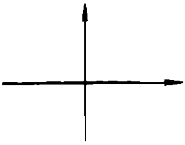
(c) If $E = 0$, we obtain the equation

$$Y^2 = 0$$

which defines *two coinciding lines* in a plane.

We may determine what type of plane curve a polynomial equation represents without making manipulations as given above. It suffices to define the signs of some expressions involving the coefficients of the polynomial equation in question.

Table 4.1 Classification of Curves of the Second Order

D	Δ	Curve	Shape
+	−	Ellipse	
	+	Imaginary ellipse	
	0	Pair of imaginary intersecting lines	
−	$\neq 0$	Hyperbola	
	0	Pair of intersecting lines	
0	$\neq 0$	Parabola	
	0	Pair of parallel lines	
		Pair of imaginary parallel lines	
		Pair of coinciding lines	

For the polynomial equation

$$ax^2 + 2bxy + cy^2 + 2dx + 2ey + g = 0$$

the criteria to determine the type of the plane curve are the signs of

$$D = \begin{vmatrix} a & b \\ b & c \end{vmatrix} \quad \text{and} \quad \Delta = \begin{vmatrix} a & b & d \\ b & c & e \\ d & e & g \end{vmatrix}.$$

The numbers D and Δ are called *invariants* since they are independent of the Cartesian coordinate system set up in a plane.

Table 4.1 shows a classification of plane curves of the second order in terms of D and Δ.

4.8 Surfaces of the Second Order

Suppose that we are given a Cartesian coordinate system in three-dimensional space. A set of points whose Cartesian coordinates x, y and z satisfy the equation

$$F(x, y, z) = 0 \qquad (*)$$

is called a *surface* in three-dimensional space. The equation (*) is called the *equation of a surface.*

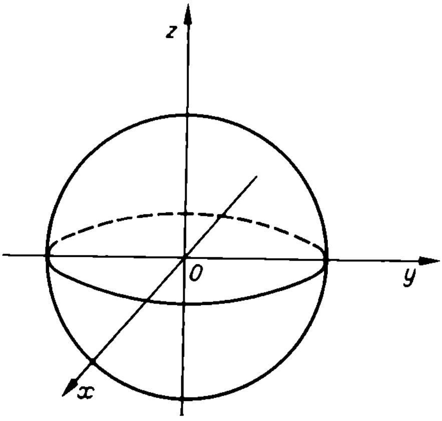


Fig. 4.36

Example. The equation

$$x^2 + y^2 + z^2 - a^2 = 0 \quad (a > 0)$$

is an equation of a sphere with radius r and centre at the point (0, 0, 0) (Fig. 4.36).

Consider the quadratic polynomial of three variables x, y and z

$$F(x, y, z) \equiv \alpha_{11}x^2 + 2\alpha_{12}xy + 2\alpha_{13}xz + \alpha_{22}y^2 + 2\alpha_{23}yz$$
$$+ \alpha_{33}z^2 + 2\alpha_{14}x + 2\alpha_{24}y + 2\alpha_{34}z + \alpha_{44} = 0,$$
$$\alpha_{11}^2 + \alpha_{12}^2 + \alpha_{13}^2 + \alpha_{22}^2 + \alpha_{23}^2 + \alpha_{33}^2 > 0.$$

The equation

$$F(x, y, z) = 0$$

is said to be the *equation of surface of the second order.*

This equation determines a rich collection of surfaces. Its investigation is a much more complicated task than that of analyzing the equation of a plane curve and requires special techniques and procedures. This will be made in Chapter 6. Here we confine ourselves to classification of surfaces in general and to standard equations of major surfaces of the second order.

4.9 Classification of Surfaces

Surfaces of revolution. We consider a curve γ (Fig. 4.37) defined in the xz-plane by the equation

$$z = f(x) \quad (x \geqslant 0).$$

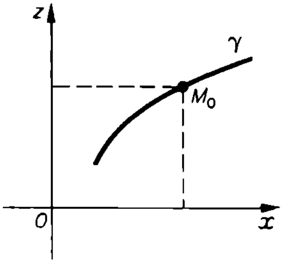


Fig. 4.37

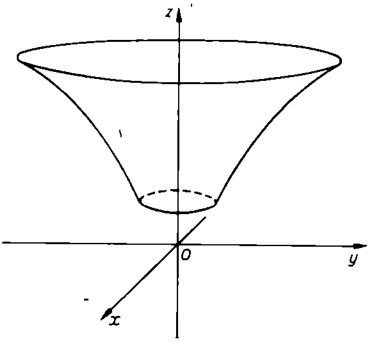


Fig. 4.38

Revolving the curve γ around the z-axis generates a surface called the *surface of revolution* (Fig. 4.38).

We find the equation of this surface, i.e., the equation to be satisfied only by points lying on the surface.

Let $M_0(x_0, 0, z_0)$ be an arbitrary point on the curve γ (Fig. 4.39). When the curve γ revolves around the z-axis the point M_0 moves in a plane which passes through M_0 and is perpendicular to the z-axis. The path of moving is a circle

$$x^2 + y^2 = x_0^2$$

with radius equal to the abscissa x_0 of M_0.

Notice that for the given curve γ $z_0 = f(x_0)$. Therefore the coordinates of any point on the circle are related as

$$z_0 = f\left(\sqrt{x^2 + y^2}\right).$$

It is easy to see that this reasoning is fully applicable to every point on the curve γ. Whence we infer that the surface of revolution is defined by the equation

$$z = f\left(\sqrt{x^2 + y^2}\right).$$

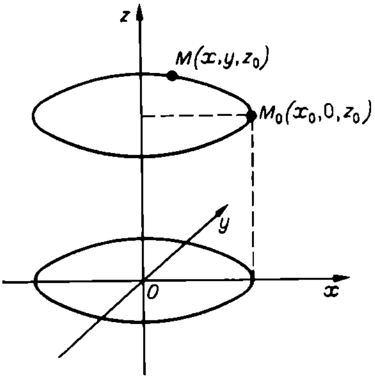


Fig. 4.39

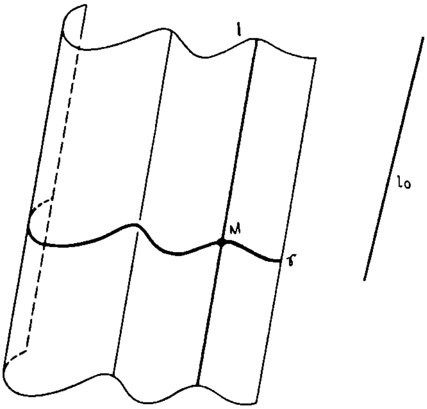


Fig. 4.40

Cylindrical surfaces. Suppose that we draw through each point of a given curve γ a line l parallel to a given line l_0. A totality of lines drawn in this way through points of the curve γ, i.e., a set of points lying on parallel lines drawn through points of the curve γ is called the *cylindrical surface*. The curve γ is called the ***directing line*** of the cylindrical surface and any line which passes through a point on the directing line and is parallel to the line l_0 is called the ***generating line*** of the cylindrical surface (Fig. 4.40).

We find the equation of the cylindrical surface.

Consider a plane π which passes through an arbitrary point O and is perpendicular to the generating line of the cylindrical surface (Fig. 4.41).

Let us set up a Cartesian coordinate system $Oxyz$ with origin at O and the z-axis perpendicular to the plane π. Then the plane π becomes a coordinate plane, i.e., the xy-plane.

It is clear that the line of intersection of the cylindrical surface and the plane π is the directing line γ_0.

Let $F(x, y) = 0$ be the equation of the directing line γ_0.

We show that the above equation may be considered the equation of a cylindrical surface.

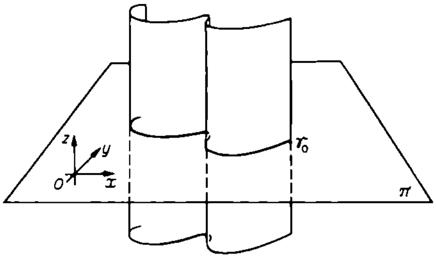


Fig. 4.41

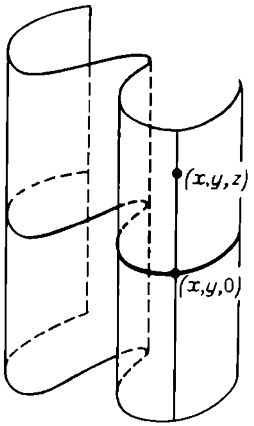


Fig. 4.42

◄ Indeed, if a point (x, y, z) lies on the cylindrical surface, then the point $(x, y, 0)$ belongs to the directing line γ_0 (Fig. 4.42). Hence the point $(x, y, 0)$ satisfies the equation

$$F(x, y) = 0.$$

On the other hand, this equation is also satisfied by the coordinates x and y of the point (x, y, z). Therefore we may regard the equation

$$F(x, y) = 0$$

as the equation of the cylindrical surface since it holds true for any point on this surface. ►

Example. Let there be given a Cartesian coordinate system $Oxyz$ in three-dimensional space (Fig. 4.43). The equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

defines a cylindrical surface called the *elliptic cylinder.*

Remark. The equation

$$F(y, z) = 0$$

defines a cylindrical surface with directing line parallel to the x-axis, and the equation

$$F(x, z) = 0$$

defines a cylindrical surface with directing line parallel to the y-axis.

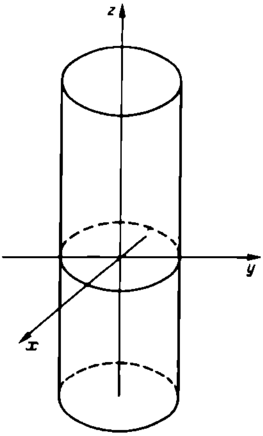


Fig. 4.43

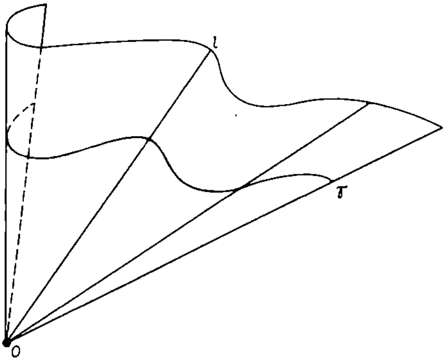


Fig. 4.44

Conic surfaces. Suppose that we are given an arbitrary curve γ and a point O outside γ. Let us draw lines through O and every point of γ. A totality of lines thus obtained, that is a set of points lying on these lines, is called the *conic surface* with the *directing line* γ and *vertex* O (Fig. 4.44). Any line passing through the vertex O and a point on the directing line γ is called the *generating line* of the conic surface.

Consider a function $F(x, y, z)$ of three variables x, y and z.

The function $F(x, y, z)$ is called a *homogeneous function of degree* q if for any $t > 0$ there holds

$$F(tx, ty, tz) = t^q F(x, y, z).$$

Let us show that given a homogeneous function $F(x, y, z)$ the equation

$$F(x, y, z) = 0$$

defines a conic surface.

◄ Indeed, let

$$F(x_0, y_0, z_0) = 0,$$

i.e., a point $M_0(x_0, y_0, z_0)$ lies on the surface defined by the equation $F(x, y, z) = 0$.

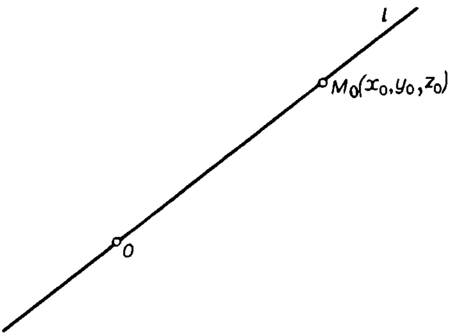


Fig. 4.45

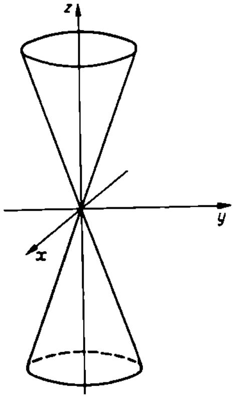


Fig. 4.46

We set $F(0, 0, 0) = 0$ and draw a line l through the points M_0 and $O(0, 0, 0)$ (Fig. 4.45). The line l is given by the parametric equations

$$x = tx_0, \quad y = ty_0, \quad z = tz_0.$$

Substituting the parametric equations into $F(x, y, z)$, we obtain

$$F(x, y, z) = F(tx_0, ty_0, tz_0) = t^q F(x_0, y_0, z_0) = 0.$$

This means that the line l belongs to the surface defined by the equation

$$F(x, y, z) = 0.$$

Hence this equation defines a conic surface. ►

Example. The function

$$F(x, y, z) = \frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2}$$

is a homogeneous function of degree 2, i.e.,

$$F(tx,\ ty,\ tz) = \frac{(tx)^2}{a^2} + \frac{(ty)^2}{b^2} - \frac{(tz)^2}{c^2}$$

$$= t^2\left(\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2}\right) = t^2F(x,\ y,\ z).$$

Whence we conclude that

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 0$$

is an equation of a conic surface (Fig. 4.46).

4.10 Standard Equations of Surfaces of the Second Order

Ellipsoids. A surface of the second order given by the standard Cartesian equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1,$$

where $a \geqslant b \geqslant c > 0$, is called the *ellipsoid.*

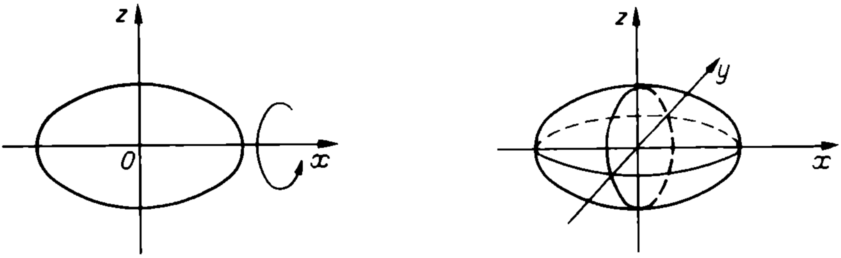


Fig. 4.47

To investigate the shape of the ellipsoid we revolve the ellipse

$$\frac{x^2}{a^2} + \frac{z^2}{c^2} = 1$$

around the z-axis (Fig. 4.47). This leads to a surface

$$\frac{x^2 + y^2}{a^2} + \frac{z^2}{c^2} = 1 \qquad (*)$$

called the *ellipsoid of revolution,* which gives an idea about the shape of the ellipsoid. It suffices to compress the ellipsoid of revolution along the y-axis with the compression coefficient $\frac{b}{a} \leqslant 1$ to get the ellipsoid given by the standard Cartesian equation. In other words, substituting $\frac{a}{b} y$ for y in (∗), we obtain the standard Cartesian equation of the ellipsoid*).

Hyperboloids. Revolving a hyperbola given by the equation

$$\frac{x^2}{a^2} - \frac{z^2}{c^2} = 1$$

around the z-axis generates a surface called the *one-sheet hyperboloid of revolution* (Fig. 4.48) which is defined by the equation

$$\frac{x^2 + y^2}{a^2} - \frac{z^2}{c^2} = 1.$$

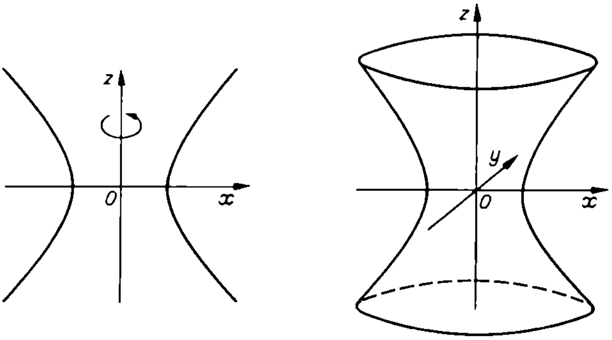


Fig. 4.48

By compressing a one-sheet hyperboloid of revolution uniformly with compression coefficient $\frac{b}{a} \leqslant 1$ along the y-axis, we obtain a *hyperboloid of one sheet* given by the standard Cartesian equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1.$$

The standard Cartesian equation of the hyperboloid of one sheet is easily obtained from the equation of the one-sheet hyperboloid of revolution by substituting $\frac{a}{b} y$ for y into the latter equation.

---

*) The ellipsoid of revolution is also obtained by uniformly compressing a sphere $x^2 + y^2 + z^2 = a^2$ with the compression coefficient $\frac{c}{a} \leqslant 1$ along the z-axis.

Revolving a conjugate hyperbola given by the equation

$$\frac{x^2}{a^2} - \frac{z^2}{c^2} = -1$$

around the z-axis generates a *two-sheet hyperboloid of revolution* (Fig. 4.49)

$$\frac{x^2 + y^2}{a^2} - \frac{z^2}{c^2} = -1.$$

As before, by compressing a two-sheet hyperboloid uniformly with compression coefficient $\frac{b}{a} \leqslant 1$ along the y-axis, we obtain a *hyperboloid of two sheets* given by the standard Cartesian equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = -1.$$

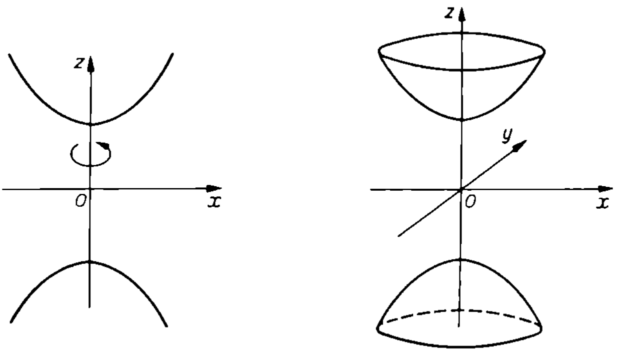


Fig. 4.49

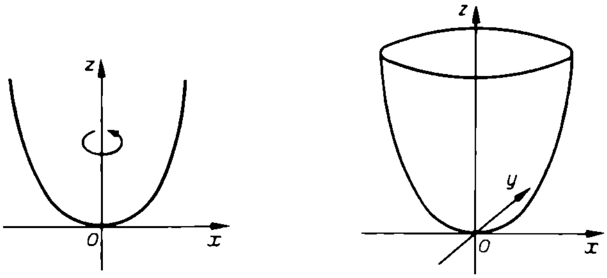


Fig. 4.50

Elliptic paraboloids. Revolving a parabola given by the equation

$$x^2 = 2pz$$

around the z-axis generates a *paraboloid of revolution* (Fig. 4.50).

$$x^2 + y^2 = 2pz.$$

By compressing a paraboloid of revolution uniformly with compression coefficient $\sqrt{\frac{q}{p}} \leqslant 1$ along the y-axis, we obtain an *elliptic paraboloid* given by the standard Cartesian equation

$$\frac{x^2}{p} + \frac{y^2}{q} = 2z.$$

It is easy to see that the standard equation of the elliptic paraboloid is obtained by substituting $\sqrt{\frac{p}{q}}\, y$ for y in the equation of the paraboloid of revolution

$$\frac{x^2 + y^2}{p} = 2z.$$

When $p < 0$, the standard Cartesian equation describes the elliptic paraboloid shown in Fig. 4.51.

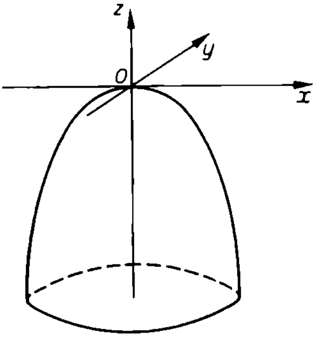


Fig. 4.51

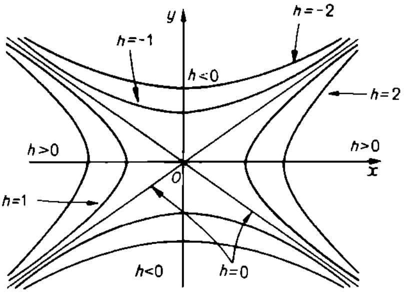


Fig. 4.52

Hyperbolic paraboloid. A surface of the second order given by a standard Cartesian equation of the form

$$\frac{x^2}{p} - \frac{y^2}{q} = 2z, \tag{*}$$

where $p > 0$ and $q > 0$, is called a *hyperbolic paraboloid.*

We shall investigate the shape of this surface by applying the following technique. Draw planes parallel to coordinate planes. These planes, called *sections,* intersect the surface in question along plane curves. Mapping the lines of intersection onto the coordinate planes, we obtain families of lines whose structures, i.e., shapes and mutual locations of the lines on the coordinate planes, enable us to make a conclusion on the shape of the surface itself.

Let us start with sections $z = h = \text{const}$ parallel to the xy-plane. Depending on the values of h, we obtain three families of intersection lines, namely

(a) a *family of hyperbolas*

$$\frac{x^2}{(\sqrt{2ph}\,)^2} - \frac{y^2}{(\sqrt{2qh}\,)^2} = 1,$$

where $h > 0$;

(b) a *family of conjugate hyperbolas*

$$\frac{x^2}{(\sqrt{-2ph}\,)^2} - \frac{y^2}{(\sqrt{-2qh})^2} = -1,$$

where $h < 0$;

(c) *two intersecting straight lines*

$$\frac{x^2}{(\sqrt{p})^2} - \frac{y^2}{(\sqrt{q})^2} = 0,$$

provided that $h = 0$.

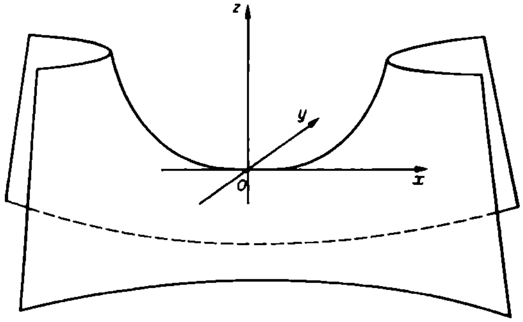


Fig. 4.53

Notice that these straight lines are asymptotes of all the hyperbolas of the families (a) and (b), i.e., they are asymptotes of hyperbolas for any value of h distinct from zero.

Mapping the intersection lines onto the xy-plane, we obtain the family of lines shown in Fig. 4.52 from which we infer that the surface in question is of a saddle shape (Fig. 4.53).

Now we cut the surface by planes $y = h$. Substituting h for y in (*), we obtain a *family of parabolas* in the xz-plane

$$x^2 = 2p\left(z + \frac{h^2}{2pq}\right),$$

as shown in Fig. 4.54.

Similarly, cutting the surface by planes $x = h$, we obtain a *family of parabolas* in the yz-plane

$$y^2 = -2q\left(z - \frac{h^2}{2pq}\right),$$

as shown in Fig. 4.55.

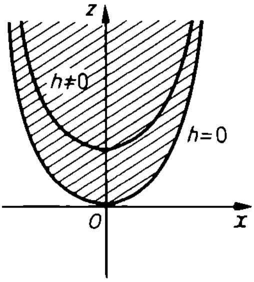


Fig. 4.54

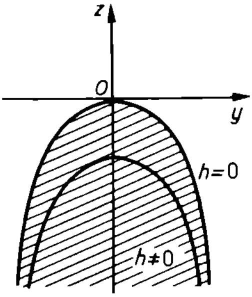


Fig. 4.55

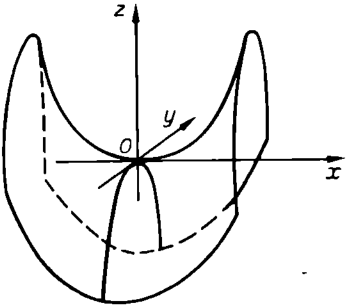


Fig. 4.56

From Figs. 4.54 and 4.55 we conclude that a hyperbolic paraboloid (Fig. 4.56) is obtained by translating a parabola $x^2 = 2pz$ along a parabola $y^2 = -2qz$ or by translating $y^2 = -2qz$ along $x^2 = 2pz$.

Remark. Intersecting a surface by a plane parallel to coordinate planes is fully applicable to the analysis of all surfaces considered above. However, revolving plane curves of the second order and further compressing surfaces thus obtained is a much easier way to investigate surfaces of the second order.

Cylinders. Recall that the shape of a cylinder (of cylindrical surface) is determined by the shape of its directing line. Here we enumerate the following kinds of cylinders we have encountered in the preceding sections. These are

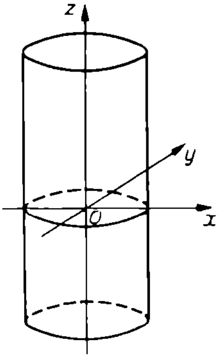


Fig. 4.57

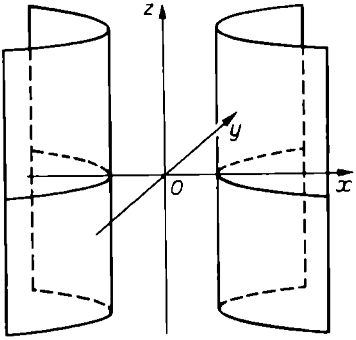


Fig. 4.58

(a) *elliptic cylinder* (Fig. 4.57)

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

(b) *hyperbolic cylinder* (Fig. 4.58)

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1,$$

(c) *parabolic cylinder* (Fig. 4.59)

$$y^2 = 2px.$$

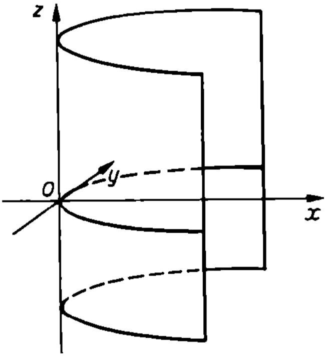


Fig. 4.59

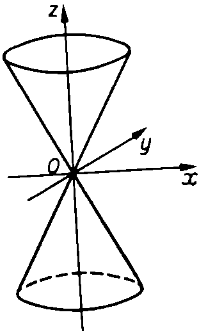


Fig. 4.60

Cones of the second order. A surface given by the standard Cartesian equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 0$$

is called a *cone of the second order.*

We may investigate the shape of this surface either by revolving two intersecting straight lines

$$\frac{x^2}{a^2} - \frac{z^2}{c^2} = 0$$

around the z-axis and further compressing the surface thus obtained or by intersecting the cone in question by planes parallel to coordinate planes. In both cases we infer that a cone of the second order is of the shape shown in Fig. 4.60.

Exercises

1. For the hyperbola $\frac{x^2}{9} - \frac{y^2}{16} = 1$ find; (a) the coordinates of foci, (b) the eccentricity, (c) the equations of asymptotes and directrices, (d) the equation of the conjugate hyperbola and its eccentricity.

2. Write down the equation of a parabola provided that the distance from the focus to the vertex is equal to 3.

3. Write down the equation of the tangent to the ellipse $\frac{x^2}{32} + \frac{y^2}{18} = 1$ at the point $M(4, 3)$.

4. Identify the types and locations of plane curves given by the equations: (a) $x^2 + 2y + 4x - 4y = 0$; (b) $6xy + 8y^2 - 12x - 26y + 11 = 0$; (c) $x^2 - 4xy + 4y^2 + 4x - 3y - 7 = 0$; (d) $xy + x + y = 0$, (e) $x^2 - 5xy + 4y^2 + x + 2y - 2 = 0$; (f) $4x^2 - 12xy + 9y^2 - 2x + 3y - 2 = 0$.

Answers

1. (a) $F_l(-5, 0)$, $F_r(5, 0)$; (b) $e = \frac{5}{3}$, (c) $y = \pm\frac{4}{3}x$, $x = \pm\frac{9}{5}$; (d) $\frac{x^2}{9} - \frac{y^2}{16} = -1$, $e = \frac{5}{4}$. **2.** $y^2 = 12x$. **3.** $3x + 4y - 24 = 0$. **4.** (a) The ellipse $\frac{X^2}{6} + \frac{Y^2}{3} = 1$ with centre at $O'(-2, 1)$ and the major axis $O'X$ parallel to the x-axis; (b) the hyperbola $\frac{X^2}{1} - \frac{Y^2}{9} = 1$ with centre at $O'(-1, 2)$ and the tangent of the angle between the axis $O'X$ and the x-axis equal to 3; (c) the parabola $Y^2 = \frac{1}{\sqrt{5}}X$ with vertex at $O'(3, 2)$, the vector of the $O'X$-axis directed to the vertex is $\{-2, -1\}$; (d) the hyperbola with centre at $O'(-1, -1)$, the asymptotes are parallel to the x- and y-axes; (e) a pair of intersecting straight lines $x - y - 1 = 0$ and $x - 4y + 2 = 0$; (f) a pair of parallel straight lines $2x - 3y + 1 = 0$ and $2x - 3y - 2 = 0$.

Chapter 5

Matrices. Determinants. Systems of Linear Equations

5.1 Matrices

Definitions. An *$m \times n$ matrix* is an array of $m \cdot n$ numbers α_{ij} ($i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$) arranged in the rectangular form

$$\mathbf{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{m1} & \alpha_{m2} & \ldots & \alpha_{mn} \end{pmatrix}. \tag{5.1}$$

The numbers α_{ij} ($i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$) are called *elements* or *entries* or *coefficients* of $\mathbf{A}$.

The horizontal n-tuple of numbers

$$\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in} \quad (i = 1, 2, \ldots, m)$$

is called the ith *row of the matrix* $\mathbf{A}$.

The vertical m-tuple of numbers

$$\begin{matrix} \alpha_{1j} \\ \alpha_{2j} \\ \vdots \\ \alpha_{mj} \end{matrix} \qquad (j = 1, 2, \ldots, n)$$

is called the *jth column of the matrix* $\mathbf{A}$.

Therefore each $m \times n$ matrix has m rows and n columns. The element α_{ij} occupies the position where the ith row and jth column meet.

The numbers i and j indicate the position of the element α_{ij} in $\mathbf{A}$ and may be thought of as the coordinates of α_{ij} in $\mathbf{A}$ (Fig. 5.1). A concise notation for matrices of the form $m \times n$ is

$$\mathbf{A} = (\alpha_{ij}).$$

A $1 \times n$ matrix is called a *row-vector* and an $m \times 1$ matrix is called a *column-vector.*

When $m = n$, the matrix

$$\mathbf{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \dots & \dots & \dots & \dots \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} \end{pmatrix}$$

is called a *square matrix* of order n.

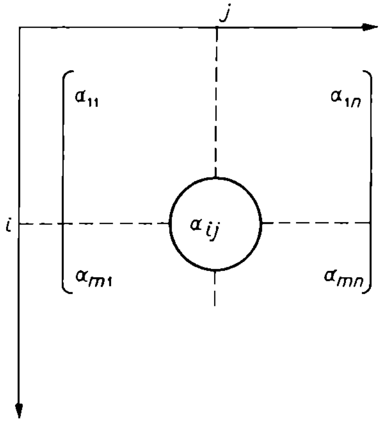


Fig. 5.1

For example, the matrix $\mathbf{A} = (\alpha_{11})$ containing a single element is a square matrix of order 1.

The n-tuple of numbers

$$\alpha_{11}, \alpha_{22}, \dots, \alpha_{nn}$$

is called the *principal diagonal* of the matrix $\mathbf{A}$.

An $m \times n$ matrix whose elements are zeros is called a *zero matrix*.

A square matrix of the form

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix} \tag{5.2}$$

is called an *identity* or *unit matrix*.

For any $m \times n$ matrix there exists a zero matrix and for any square matrix of a given order n there exists a unit matrix.

We shall denote the set of all matrices of the type $m \times n$ by $\mathbb{R}^{m \times n}$ with the understanding that we are concerned with matrices whose elements are real numbers. Since a set of real numbers is conventionally denoted by $\mathbb{R}$, the notation $\mathbb{R}^{m \times n}$ signifies that we consider a set of all $m \times n$ matrices of real numbers. In Chapter 26 we shall consider a set of $m \times n$ matrices of complex numbers denoted by $\mathbb{C}^{m \times n}$ because a set of complex numbers is conventionally denoted by $\mathbb{C}$.

To signify that matrix (5.1) is of the type $m \times n$ we shall write

$$\mathbf{A} = (\alpha_{ij}) \in \mathbb{R}^{m \times n}.$$

The matrices $\mathbf{A} = (\alpha_{ij})$ and $\mathbf{B} = (\beta_{ij})$ are called *equal* if they are of the same type and their elements occupying identical positions coincide, i.e., if

$$\mathbf{A} \in \mathbb{R}^{m \times n}, \quad \mathbf{B} \in \mathbb{R}^{m \times n}$$

and

$$\alpha_{ij} = \beta_{ij} \quad (i = 1, 2, \ldots, m; \; j = 1, 2, \ldots, n).$$

In symbols, we write $\mathbf{A} = \mathbf{B}$.

Now we shall turn our attention to arithmetic operations on matrices.

Addition of matrices. Let $\mathbf{A}$ and $\mathbf{B}$ be two matrices of type $m \times n$, that is,

$$\mathbf{A} = (\alpha_{ij}) \in \mathbb{R}^{m \times n} \quad \text{and} \quad \mathbf{B} = (\beta_{ij}) \in \mathbb{R}^{m \times n}.$$

The *sum* of matrices $\mathbf{A}$ and $\mathbf{B}$ is the matrix $\mathbf{C} = (\gamma_{ij}) \in \mathbb{R}^{m \times n}$ whose elements are

$$\gamma_{ij} = \alpha_{ij} + \beta_{ij} \quad (i = 1, 2, \ldots, m; \; j = 1, 2, \ldots, n). \tag{5.3}$$

In symbols, we write $\mathbf{C} = \mathbf{A} + \mathbf{B}$.

Multiplication of a matrix by a scalar. The *product* of a matrix $\mathbf{A} = (\alpha_{ij}) \in \mathbb{R}^{m \times n}$ by a scalar λ is the matrix $\mathbf{B} = (\beta_{ij}) \in \mathbb{R}^{m \times n}$ whose elements are

$$\beta_{ij} = \lambda\alpha_{ij} \quad (i = 1, 2, \ldots, m; \; j = 1, 2, \ldots, n). \tag{5.4}$$

In symbols, we write $\mathbf{B} = \lambda\mathbf{A}$.

By way of illustration we show how these operations are performed by using notation (5.1):

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{m1} & \alpha_{m2} & \ldots & \alpha_{mn} \end{pmatrix} + \begin{pmatrix} \beta_{11} & \beta_{12} & \ldots & b_{1n} \\ \beta_{21} & \beta_{22} & \ldots & \beta_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \beta_{m1} & \beta_{m2} & \ldots & \beta_{mn} \end{pmatrix}$$

$$= \begin{pmatrix} \alpha_{11} + \beta_{11} & \alpha_{12} + \beta_{12} & \ldots & \alpha_{1n} + \beta_{1n} \\ \alpha_{21} + \beta_{21} & \alpha_{22} + \beta_{22} & \ldots & \alpha_{2n} + \beta_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{m1} + \beta_{m1} & \alpha_{m2} + \beta_{m2} & \ldots & \alpha_{mn} + \beta_{mn} \end{pmatrix},$$

$$\lambda\begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \hdotsfor{4} \\ \alpha_{m1} & \alpha_{m2} & \dots & \alpha_{mn} \end{pmatrix} = \begin{pmatrix} \lambda\alpha_{11} & \lambda\alpha_{12} & \dots & \lambda\alpha_{1n} \\ \lambda\alpha_{21} & \lambda\alpha_{22} & \dots & \lambda\alpha_{2n} \\ \hdotsfor{4} \\ \lambda\alpha_{m1} & \lambda\alpha_{m2} & \dots & \lambda\alpha_{mn} \end{pmatrix}.$$

Multiplication of matrices. Let $\mathbf{A} = (\alpha_{ik})$ and $\mathbf{B} = (\beta_{kj})$ be two square matrices of order n. The *product* of $\mathbf{A}$ and $\mathbf{B}$ is the matrix $\mathbf{C} = (\gamma_{ij}) \in \mathbb{R}^{n \times n}$ whose elements are

$$\gamma_{ij} = \alpha_{i1}\beta_{1j} + \alpha_{i2}\beta_{2j} + \dots + \alpha_{in}\beta_{nj} \qquad (i, j = 1, 2, \dots, n). \tag{5.5}$$

In symbols, we write $\mathbf{C} = \mathbf{AB}$.

By way of illustration we write this operation as

$$\begin{array}{c@{}c} \begin{matrix} \\ \\ \\ i \\ \\ \\ \end{matrix} & \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1j} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2j} & \dots & \alpha_{2n} \\ \hdotsfor{6} \\ \alpha_{i1} & \alpha_{i2} & \dots & \alpha_{ij} & \dots & \alpha_{in} \\ \hdotsfor{6} \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nj} & \dots & \alpha_{nn} \end{pmatrix} \end{array}$$

$$\times \begin{array}{c} j \\ \begin{pmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1j} & \dots & \beta_{1n} \\ \beta_{21} & \beta_{22} & \dots & \beta_{2j} & \dots & \beta_{2n} \\ \hdotsfor{6} \\ \beta_{i1} & \beta_{i2} & \dots & \beta_{ij} & \dots & \beta_{in} \\ \hdotsfor{6} \\ \beta_{n1} & \beta_{n2} & \dots & \beta_{nj} & \dots & \beta_{nn} \end{pmatrix} \end{array}$$

$$= \begin{array}{c@{}c} j & \\ \begin{pmatrix} \gamma_{11} & \gamma_{12} & \dots & \gamma_{1j} & \dots & \gamma_{1n} \\ \gamma_{21} & \gamma_{22} & \dots & \gamma_{2j} & \dots & \gamma_{2n} \\ \hdotsfor{6} \\ \gamma_{i1} & \gamma_{i2} & \dots & \gamma_{ij} & \dots & \gamma_{in} \\ \hdotsfor{6} \\ \gamma_{n1} & \gamma_{n2} & \dots & \gamma_{nj} & \dots & \gamma_{nn} \end{pmatrix} & \begin{matrix} \\ \\ \\ i. \\ \\ \\ \end{matrix} \end{array}$$

Notice that the sum of n elements (numbers)

$$\alpha_1 + \alpha_2 + \dots + \alpha_i + \dots + \alpha_n$$

is conventionally written in the form

$$\sum_{i=1}^{n} \alpha_i.$$

Then the operation of multiplication of two matrices may be conveniently presented as

$$\gamma_{ij} = \sum_{k=1}^{n} \alpha_{ik}\beta_{kj} \quad (i, j = 1, 2, \ldots, n).$$

In general $\mathbf{AB} \neq \mathbf{BA}$.

Examples. (1) Let

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

Then

$$\mathbf{AB} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{BA} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

It is not hard to find similar examples for matrices of any order.

(2) Let $\mathbf{A}$ be a square matrix of order 3, i.e.,

$$\mathbf{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix}.$$

◄ We show that pre-multiplying the matrix $\mathbf{A}$ by the matrix

$$\mathbf{P}_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

interchanges the second and third rows in $\mathbf{A}$.

We have

$$\mathbf{P}_{23}\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix}$$

$$= \begin{pmatrix} 1\cdot\alpha_{11} + 0\cdot\alpha_{21} + 0\cdot\alpha_{31} & 1\cdot\alpha_{12} + 0\cdot\alpha_{22} + 0\cdot\alpha_{32} & 1\cdot\alpha_{13} + 0\cdot\alpha_{23} + 0\cdot\alpha_{33} \\ 0\cdot\alpha_{11} + 0\cdot\alpha_{21} + 1\cdot\alpha_{31} & 0\cdot\alpha_{12} + 0\cdot\alpha_{22} + 1\cdot\alpha_{32} & 0\cdot\alpha_{13} + 0\cdot\alpha_{23} + 1\cdot\alpha_{33} \\ 0\cdot\alpha_{11} + 1\cdot\alpha_{21} + 0\cdot\alpha_{31} & 0\cdot\alpha_{12} + 1\cdot\alpha_{22} + 0\cdot\alpha_{32} & 0\cdot\alpha_{13} + 1\cdot\alpha_{23} + 0\cdot\alpha_{33} \end{pmatrix}$$

$$= \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \end{pmatrix}. \blacktriangleright$$

Similarly, it is easy to verify that post-multiplying the matrix $\mathbf{A}$ by $\mathbf{P}_{23}$ interchanges the second and third columns in $\mathbf{A}$.

(3) For any matrix **A** there holds

$$\mathbf{I}\cdot\mathbf{A} = \mathbf{A}\cdot\mathbf{I} = \mathbf{A}, \tag{5.6}$$

where **I** is an identity matrix.

◄ Let **A** be a square matrix of order 3

$$\mathbf{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix}.$$

Then

$$\mathbf{A}\cdot\mathbf{I} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix}\cdot\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \alpha_{11}\cdot 1 + \alpha_{12}\cdot 0 + \alpha_{13}\cdot 0 & \alpha_{11}\cdot 0 + \alpha_{12}\cdot 1 + \alpha_{13}\cdot 0 & \alpha_{11}\cdot 0 + \alpha_{12}\cdot 0 + \alpha_{13}\cdot 1 \\ \alpha_{21}\cdot 1 + \alpha_{22}\cdot 0 + \alpha_{23}\cdot 0 & \alpha_{21}\cdot 0 + \alpha_{22}\cdot 1 + \alpha_{23}\cdot 0 & \alpha_{21}\cdot 0 + \alpha_{22}\cdot 0 + \alpha_{23}\cdot 1 \\ \alpha_{31}\cdot 1 + \alpha_{32}\cdot 0 + \alpha_{33}\cdot 0 & \alpha_{31}\cdot 0 + \alpha_{32}\cdot 1 + \alpha_{33}\cdot 0 & \alpha_{31}\cdot 0 + \alpha_{32}\cdot 0 + \alpha_{33}\cdot 1 \end{pmatrix}$$

$$= \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix} = \mathbf{A}.$$

It is easy to verify in the same way that

$\mathbf{I}\cdot\mathbf{A} = \mathbf{A}$. ►

Relation (5.6) explains why the matrix **I** is called the identity matrix.

The operation of matrix multiplication obeys associative and distributive laws, namely if **A**, **B**, **C** and **D** are square matrices of order n, then

$$\begin{aligned} (\mathbf{AB})\mathbf{C} &= \mathbf{A}(\mathbf{BC}), \\ \mathbf{A}(\mathbf{B} + \mathbf{C}) &= \mathbf{AB} + \mathbf{AC}, \\ (\mathbf{B} + \mathbf{C})\mathbf{D} &= \mathbf{BD} + \mathbf{CD}. \end{aligned}$$

◄ By way of illustration we show that $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$. It is clear that the three matrices **AB**, **AC** and $\mathbf{A}(\mathbf{B} + \mathbf{C})$ are of the same order n. Their elements in the (i, j)th position are, respectively,

$$\sum_{k=1}^{n} \alpha_{ik}\beta_{kj},$$

$$\sum_{k=1}^{n} \alpha_{ik}\gamma_{kj}, \qquad (i, j = 1, 2, \ldots, n).$$

$$\sum_{k=1}^{n} \alpha_{ik}(\beta_{kj} + \gamma_{kj}),$$

Then we obtain

$$\sum_{k=1}^{n} \alpha_{ik}(\beta_{kj} + \gamma_{kj}) = \sum_{k=1}^{n} \alpha_{ik}\beta_{kj} + \sum_{k=1}^{n} \alpha_{ik}\gamma_{kj}$$

and this proves the property in question.

Similar reasoning is applicable to the other two identities. ►

Remark. The operation of matrix multiplication may also be defined for rectangular matrices.

◄ Let there be given matrices $\mathbf{A} = (\alpha_{ik}) \in \mathbb{R}^{m \times n}$ and $\mathbf{B} = (\beta_{kj}) \in \mathbb{R}^{n \times l}$. Then the elements of the matrix $\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m \times l}$ are

$$\gamma_{ij} = \sum_{k=1}^{n} \alpha_{ik}\beta_{kj} \quad (i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, l). \tag{5.7}$$

Therefore the product of two rectangular matrices may be defined only when the number of columns in the first factor is equal to the number of rows in the second factor (Fig. 5.2) as is easily seen from (5.7). ►

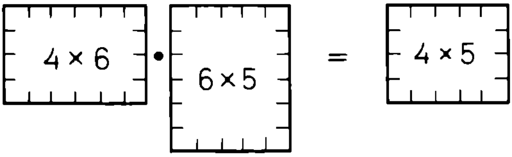


Fig. 5.2

Products of rectangular matrices obey the laws (5.6) provided that the corresponding operations of matrix multiplication make sense.

Example. Compute the product of the matrix

$$\mathbf{A} = \begin{pmatrix} 9 & 5 \\ 1 & 9 \\ 8 & 6 \end{pmatrix} \text{ by the matrix } \mathbf{B} = \begin{pmatrix} 1 & 0 & 1 & 2 \\ 1 & 9 & 4 & 2 \end{pmatrix}.$$

◄ Observe that the number of columns in **A** is equal to the number of rows in **B**. This means that we may multiply **A** by **B**. Computing the product, we obtain

$$\begin{aligned} \mathbf{AB} &= \begin{pmatrix} 9 & 5 \\ 1 & 9 \\ 8 & 6 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 1 & 2 \\ 1 & 9 & 4 & 2 \end{pmatrix} \\ &= \begin{pmatrix} 9\cdot1 + 5\cdot1 & 9\cdot0 + 5\cdot9 & 9\cdot1 + 5\cdot4 & 9\cdot2 + 5\cdot2 \\ 1\cdot1 + 9\cdot1 & 1\cdot0 + 9\cdot9 & 1\cdot1 + 9\cdot4 & 1\cdot2 + 9\cdot2 \\ 8\cdot1 + 6\cdot1 & 8\cdot0 + 6\cdot9 & 8\cdot1 + 6\cdot4 & 8\cdot2 + 6\cdot2 \end{pmatrix} \\ &= \begin{pmatrix} 14 & 45 & 29 & 28 \\ 10 & 81 & 37 & 20 \\ 14 & 54 & 32 & 28 \end{pmatrix}. \end{aligned}$$

►

Summing up elements of a matrix. In applications we often need to compute the sum of elements of a rectangular matrix, say the matrix **A**

$$\mathbf{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \dots & \dots & \dots & \dots \\ \alpha_{m1} & \alpha_{m2} & \dots & \alpha_{mn} \end{pmatrix}.$$

We may compute the sum **H** of elements of **A** either by computing column sums

$$\sum_{i=1}^{m} \alpha_{i1}, \quad \sum_{i=1}^{m} \alpha_{i2}, \quad \dots, \quad \sum_{i=1}^{m} \alpha_{in}$$

and adding them together so that

$$\mathbf{H} = \sum_{i=1}^{m} \alpha_{i1} + \sum_{i=1}^{m} \alpha_{i2} + \dots + \sum_{i=1}^{m} \alpha_{in} = \sum_{j=1}^{n} \left(\sum_{i=1}^{m} \alpha_{ij} \right),$$

or by finding row sums

$$\sum_{j=1}^{n} \alpha_{1j}, \quad \sum_{j=1}^{n} \alpha_{2j}, \quad \dots, \quad \sum_{j=1}^{n} \alpha_{mj}$$

and adding them together so that

$$\mathbf{H} = \sum_{j=1}^{n} \alpha_{1j} + \sum_{j=1}^{n} \alpha_{2j} + \dots + \sum_{j=1}^{n} \alpha_{mj} = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \alpha_{ij} \right).$$

Whence we obtain

$$\sum_{j=1}^{n} \sum_{i=1}^{m} \alpha_{ij} = \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_{ij}.$$

Transposition of matrices. The matrix

$$\begin{pmatrix} \alpha_{11} & \alpha_{21} & \dots & \alpha_{m1} \\ \alpha_{12} & \alpha_{22} & \dots & \alpha_{m2} \\ \dots & \dots & \dots & \dots \\ \alpha_{1n} & \alpha_{2n} & \dots & \alpha_{mn} \end{pmatrix}$$

is called the *transpose* of the matrix

$$\mathbf{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \dots & \dots & \dots & \dots \\ \alpha_{m1} & \alpha_{m2} & \dots & \alpha_{mn} \end{pmatrix}. \tag{5.1}$$

In symbols, we write **A**′ to denote the transpose of **A**.

Example. By definition the transpose of

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix}$$

is

$$\mathbf{A}' = \begin{pmatrix} 1 & 5 \\ 2 & 6 \\ 3 & 7 \\ 4 & 8 \end{pmatrix}.$$

It is important to observe that the element of $\mathbf{A}'$ in the (i, j)th position coincides with the element of $\mathbf{A}$ in the (j, i)th position. The operation of transposition interchanges rows and columns of the matrix $\mathbf{A}$ so that rows in $\mathbf{A}$ become columns in $\mathbf{A}'$ and columns in $\mathbf{A}$ become rows in $\mathbf{A}'$. Therefore if the matrix $\mathbf{A}$ has m rows and n columns, the transpose $\mathbf{A}'$ of $\mathbf{A}$ contains n rows and m columns (Fig. 5.3).

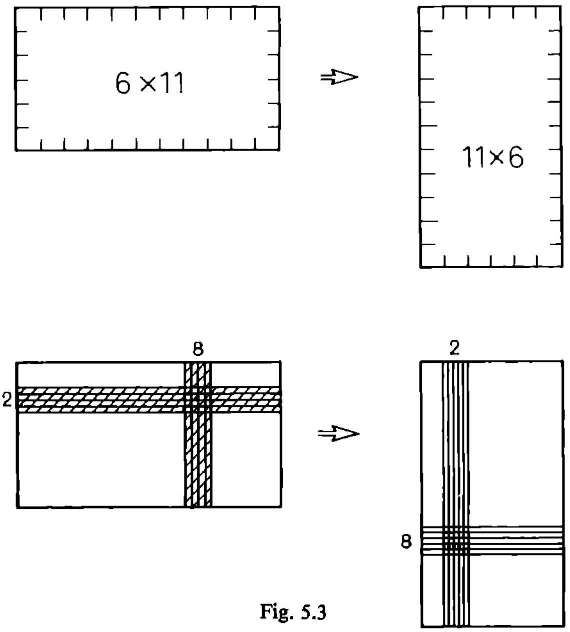


Fig. 5.3

The operation of transposition satisfies the following conditions:

(a) $(\mathbf{A}')' = \mathbf{A}$,
(b) $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$,
(c) $(\lambda\mathbf{A})' = \lambda\mathbf{A}'$,
(d) $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.

Linear dependence of row-vectors. We consider the operations of matrix addition and multiplication of a matrix by a scalar over a set of 1 × n matrices, i.e., row-vectors.

Let

$$\mathbf{a} = (\alpha_1, \alpha_2, \ldots, \alpha_n) \in \mathbb{R}^{1\times n},$$
$$\mathbf{b} = (\beta_1, \beta_2, \ldots, \beta_n) \in \mathbb{R}^{1\times n}.$$

Using (5.3) and (5.4), we obtain

$$\mathbf{a} + \mathbf{b} = (\alpha_1 + \beta_1, \alpha_2 + \beta_2, \ldots, \alpha_n + \beta_n) \tag{5.8}$$

and

$$\lambda\mathbf{a} = (\lambda\alpha_1, \lambda\alpha_2, \ldots, \lambda\alpha_n). \tag{5.9}$$

It is easy to verify that operations (5.8) and (5.9) satisfy the following conditions:

$$\begin{aligned}
&\text{(a) } \mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a},\\
&\text{(b) } (\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c}),\\
&\text{(c) } \mathbf{a} + \mathbf{0} = \mathbf{0} + \mathbf{a} = \mathbf{a},\\
&\text{(d) } \lambda(\mathbf{a} + \mathbf{b}) = \lambda\mathbf{a} + \lambda\mathbf{b},\\
&\text{(e) } (\lambda + \mu)\mathbf{a} = \lambda\mathbf{a} + \mu\mathbf{a},\\
&\text{(f) } \lambda(\mu\mathbf{a}) = (\lambda\mu)\mathbf{a},\\
&\text{(g) } 1\cdot\mathbf{a} = \mathbf{a},
\end{aligned} \tag{5.10}$$

the equation $\mathbf{a} + \mathbf{x} = \mathbf{0}$ being uniquely soluble for any row-vector $\mathbf{a}$. In (5.10) λ and μ are arbitrary scalars, $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ and $\mathbf{x}$ are row-vectors (1 × n matrices), $\mathbf{0}$ is a zero row-vector (a zero 1 × n matrix). We shall see in Chap. 6 that conditions (5.10) define a linear vector space over a set of row-vectors.

Now we shall introduce an important notion of linear dependence of row-vectors.

Let $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ be m row-vectors. A relation of the form

$$\mathbf{b} = \lambda_1\mathbf{a}_1 + \lambda_2\mathbf{a}_2 + \ldots + \lambda_m\mathbf{a}_m \tag{5.11}$$

is called a *linear combination* of row-vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ with scalars $\lambda_1, \lambda_2, \ldots, \lambda_m$.

The linear combination (5.11) is called *nontrivial* if at least one of λ_1, $\lambda_2, \ldots, \lambda_m$ is distinct from zero and *trivial* if $\lambda_1 = \lambda_2 = \ldots = \lambda_m = 0$. It is easy to see that in the latter case $\mathbf{b}$ is a zero row-vector.

Row-vectors are called *linearly dependent* if there exists their nontrivial linear combination identical to a zero row-vector, and *linearly independent* if a zero row-vector corresponds to their trivial linear combination only.

We show that for linearly dependent row-vectors one of them may be expressed as a linear combination of the others.

◄ Let $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ be linearly dependent row-vectors. This means that there exist scalars $\lambda_1, \lambda_2, \ldots, \lambda_m$, not all zero, such that

$$\lambda_1\mathbf{a}_1 + \lambda_2\mathbf{a}_2 + \ldots + \lambda_{m-1}\mathbf{a}_{m-1} + \lambda_m\mathbf{a}_m = \mathbf{0}.$$

Without a loss of generality we may set $\lambda_m \neq 0$. Transposing the first $m - 1$ summands to the right, we obtain

$$\lambda_m\mathbf{a}_m = -\lambda_1\mathbf{a}_1 - \lambda_2\mathbf{a}_2 - \ldots - \lambda_{m-1}\mathbf{a}_{m-1}.$$

Then dividing the above expression by $\lambda_m \neq 0$, we arrive at

$$\mathbf{a}_m = -\frac{\lambda_1}{\lambda_m}\mathbf{a}_1 - \frac{\lambda_2}{\lambda_m}\mathbf{a}_2 - \ldots - \frac{\lambda_{m-1}}{\lambda_m}\mathbf{a}_{m-1}.$$

Whence we conclude that $\mathbf{a}_m$ is a linear combination of the other row-vectors.

The converse is also true, namely, if one of the row-vectors is a linear combination of the other $(m - 1)$ row-vectors, i.e., if

$$\mathbf{a}_m = \mu_1\mathbf{a}_1 + \mu_2\mathbf{a}_2 + \ldots + \mu_{m-1}\mathbf{a}_{m-1},$$

then the nontrivial linear combination

$$\mu_1\mathbf{a}_1 + \mu_2\mathbf{a}_2 + \ldots + \mu_{m-1}\mathbf{a}_{m-1} + (-1)\mathbf{a}_m$$

of the row-vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{m-1}, \mathbf{a}_m$ is equal to a zero row-vector. Hence the row-vectors are linearly dependent. ►

Similar property of linear dependence is easily derived for the set $\mathbb{R}^{m\times 1}$ of column-vectors, that is, for the set of all $m \times 1$ matrices.

Elementary operations on matrices. Let $\mathbf{A}$ and $\tilde{\mathbf{A}}$ be two arbitrary $m \times n$ matrices and let

$$\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k, \ldots, \mathbf{a}_l, \ldots, \mathbf{a}_m$$

be rows in $\mathbf{A}$.

The matrix $\tilde{\mathbf{A}}$ is said to be obtained from the matrix $\mathbf{A}$ by

(a) *interchanging two rows in* $\mathbf{A}$ if $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_l, \ldots, \mathbf{a}_k, \ldots, \mathbf{a}_m$ are rows of $\tilde{\mathbf{A}}$,

(b) *multiplying one particular row of* $\mathbf{A}$ *by a nonzero scalar* β if $\mathbf{a}_1, \mathbf{a}_2, \ldots, \beta\mathbf{a}_k, \ldots, \mathbf{a}_l, \ldots, \mathbf{a}_m$ are rows of $\tilde{\mathbf{A}}$,

(c) *adding one row of* $\mathbf{A}$ *multiplied by a scalar* γ *to another row* if $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k, \ldots, \mathbf{a}_l + \gamma\mathbf{a}_k, \ldots, \mathbf{a}_m$ are rows of $\tilde{\mathbf{A}}$.

The operations (a)-(c) are called the *elementary row operations.*

Elementary column operations can be defined similarly.

Example. The matrix

$$\begin{pmatrix} 1 & 0 & 9 \\ 0 & 6 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

is obtained from the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 9 \\ 0 & 1 & 0 \\ 0 & 6 & 1 \end{pmatrix}$$

by interchanging the second and third rows, and the matrix

$$\begin{pmatrix} 0 & 1 & 9 \\ 1 & 0 & 0 \\ 6 & 0 & 1 \end{pmatrix}$$

is obtained from **A** by interchanging the first and second columns.

Adding the third row of **A** multiplied by -2 to the first row of **A**, we obtain the matrix

$$\begin{pmatrix} 1 & -12 & 7 \\ 0 & 1 & 0 \\ 0 & 6 & 1 \end{pmatrix}.$$

Remark. It is easy to see that when the matrix $\tilde{\mathbf{A}}$ is obtained by applying either of the elementary row operations to the matrix **A**, the transition from $\tilde{\mathbf{A}}$ to **A** is achieved by the same row operation, that is, either by interchanging the kth and lth rows or by multiplying the kth row by the scalar $1/\beta$ or by adding the kth row multiplied by $-\gamma$ to the lth row.

Now we shall turn our attention to the procedure of changing from an arbitrary matrix to a matrix of simpler form by a finite sequence of elementary row operations.

Let $\mathbf{A} = (\alpha_{ij}) \in \mathbb{R}^{m \times n}$ be a nonzero matrix.

Step 1. Since **A** is a nonzero matrix there exists at least one nonzero element in **A**. Consequently, there exists at least one nonzero row in **A**. We choose a nonzero row such that its first nonzero element occurs in a column with the smallest number $k_1 \geqslant 1$. Interchanging this row and the first row of **A**, we reduce **A** to the matrix

$$\begin{pmatrix} 0 & \dots & 0 & \alpha_{1k_1}^{(1)} & \dots & \alpha_{1n}^{(1)} \\ 0 & \dots & 0 & \alpha_{2k_1} & \dots & \alpha_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \alpha_{mk_1} & \dots & \alpha_{mn} \end{pmatrix}, \tag{5.12}$$

where $\alpha_{1k_1}^{(1)} \neq 0$.

Adding scalar multiples of the first row of (5.12) by $-\dfrac{\alpha_{ik_1}}{\alpha^{(1)}_{1k_1}}$ $(i = 2, 3, \ldots, m)$ to the corresponding rows, we obtain a matrix of the form

$$\begin{pmatrix} 0 & \ldots & 0 & \alpha^{(1)}_{1k_1} & \ldots & \alpha^{(1)}_{1n} \\ 0 & \ldots & 0 & 0 & \ldots & \alpha^{(1)}_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \ldots & 0 & 0 & \ldots & \alpha^{(1)}_{mn} \end{pmatrix}. \tag{5.13}$$

Observe that the elements in the k_1th row are zeros with the exception of the element $\alpha^{(1)}_{1k_1}$.

It may happen that rows of (5.13) are zero ones with the exception of the first row. In this case the procedure terminates. If this is not the case, i.e., if there exist nonzero rows in addition to the first row, the procedure goes on.

Step 2. Similar to Step 1 we choose a row such that its first nonzero element occurs in a column with the smallest number k_2 $(k_1 < k_2)$. Then interchanging this row and the second row of (5.13), we obtain

$$\begin{pmatrix} 0 & \ldots & 0 & \alpha^{(1)}_{1k_1} & \ldots & \alpha^{(1)}_{1k_2-1} & \alpha^{(1)}_{1k_2} & \ldots & \alpha^{(1)}_{1n} \\ 0 & \ldots & 0 & 0 & \ldots & 0 & \alpha^{(2)}_{2k_2} & \ldots & \alpha^{(2)}_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \ldots & 0 & 0 & \ldots & 0 & \alpha^{(1)}_{mk_2} & \ldots & \alpha^{(1)}_{mn} \end{pmatrix}, \tag{5.14}$$

where $\alpha^{(2)}_{2k_2} \neq 0$.

Adding scalar multiples of the second row by $-\dfrac{\alpha^{(1)}_{ik_2}}{\alpha^{(2)}_{2k_2}}$ $(i = 3, 4, \ldots, m)$ to the corresponding rows, we reduce (5.14) to a matrix such that its first row is identical to that of the first row in (5.13) and its elements in the k_2th row are zeros with the exception of two elements from the top.

The procedure terminates if the matrix thus obtained contains no nonzero rows but the first and second rows, and goes on otherwise. It is important to observe that for the procedure discussed the total number of steps to reduce a given matrix to a simpler form does not exceed the number $\min(m, n)$. This means that the procedure terminates in a finite number of successive steps and we arrive at a matrix of the form

$$\left(\begin{array}{cccccccccccc} 0 & \ldots & 0 & \multicolumn{1}{|c}{\alpha^{(1)}_{1k_1}} & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \cline{4-6} 0 & \ldots & 0 & 0 & \ldots & \multicolumn{1}{c|}{0} & \alpha^{(2)}_{2k_2} & \ldots & \ldots & \ldots & \ldots & \ldots \\ \cline{7-9} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \ldots & 0 & 0 & \ldots & 0 & 0 & \ldots & \multicolumn{1}{c|}{0} & \alpha^{(r)}_{rk_r} & \ldots & \\ \cline{10-12} 0 & \ldots & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \ldots & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \end{array}\right), \tag{5.15}$$

where $k_1 < k_2 < \ldots < k_r$ and

$$\alpha^{(1)}_{1k_1} \neq 0, \quad \alpha^{(2)}_{2k_2} \neq 0, \quad \ldots, \quad \alpha^{(r)}_{rk_r} \neq 0.$$

The matrix (5.15) is called a *matrix of the schematic form.*

Notice that the procedure of reducing a matrix to the schematic form involves a sequence of elementary row operations (a) and (c). In other words, we may state the following theorem.

Theorem 5.1. *The transition from any arbitrary matrix to a matrix of schematic form is achieved by a sequence of elementary row operations.*

Examples. **(1)** Reduce the matrix

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & -2 & 3 & -4 & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

to a matrix of schematic form.

◄ Interchanging the first and fourth rows of **A**, we obtain

$$\mathbf{A}_1 = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & -2 & 3 & -4 & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Interchanging the third and fourth rows of $\mathbf{A}_1$, we obtain

$$\mathbf{A}_2 = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & -2 & 3 & -4 & 5 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The matrix $\mathbf{A}_2$ is of schematic form. ►

(2) Reduce the matrix

$$\mathbf{A} = \begin{pmatrix} 3 & -1 & 3 & 2 & 5 \\ 5 & -3 & 2 & 3 & 4 \\ 1 & -3 & -5 & 0 & -7 \\ 7 & -5 & 1 & 4 & 1 \end{pmatrix}$$

to a matrix of schematic form.

◄ Interchanging the first and third rows, we obtain

$$\mathbf{A}_1 = \begin{pmatrix} 1 & -3 & -5 & 0 & -7 \\ 5 & -3 & 2 & 3 & 4 \\ 3 & -1 & 3 & 2 & 5 \\ 7 & -5 & 1 & 4 & 1 \end{pmatrix}.$$

Step 1. Subtracting the first row of $\mathbf{A}_1$ multiplied by numbers 5, 3 and 7 from the second, third and fourth rows, respectively, we obtain

$$\mathbf{A}_2 = \begin{pmatrix} 1 & -3 & -5 & 0 & 7 \\ 0 & 12 & 27 & 3 & 39 \\ 0 & 8 & 18 & 2 & 26 \\ 0 & 16 & 36 & 4 & 50 \end{pmatrix}.$$

Step 2. To simplify computations we apply the elementary operation (b), namely, we multiply the second row by 1/3, the third row by 1/2 and the fourth row by 1/2. Then from $\mathbf{A}_2$ we obtain

$$\mathbf{A}_3 = \begin{pmatrix} 1 & -3 & 5 & 0 & -7 \\ 0 & 4 & 9 & 1 & 13 \\ 0 & 4 & 9 & 1 & 13 \\ 0 & 8 & 18 & 2 & 25 \end{pmatrix}.$$

Subtracting the second row of $\mathbf{A}_3$ multiplied by 1 and 2 from the third and fourth rows, respectively, we arrive at

$$\mathbf{A}_4 = \begin{pmatrix} 1 & -3 & 5 & 0 & -7 \\ 0 & 4 & 9 & 1 & 13 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

Step 3. Observe that the third row of $\mathbf{A}_4$ is a zero one. Then interchanging the third and fourth rows, we obtain

$$\mathbf{A}_5 = \begin{pmatrix} 1 & -3 & 5 & 0 & 7 \\ 0 & 4 & 9 & 1 & 13 \\ 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The matrix $\mathbf{A}_5$ is of schematic form. ►

By a sequence of elementary column operations a matrix of the form (5.15) is reduced to the form

$$\begin{array}{cc}
 & \begin{array}{cccccc} & & & r & & \end{array} \\
r & \left(\begin{array}{cccccc|c}
1 & 0 & \dots & 0 & \dots & 0 & \\
0 & 1 & \dots & 0 & \dots & 0 & \\
\multicolumn{6}{c}{\dots\dots\dots\dots\dots\dots\dots} & 0 \\
0 & 0 & \dots & 1 & \dots & 0 & \\
\multicolumn{6}{c}{\dots\dots\dots\dots\dots\dots\dots} & \\
0 & 0 & \dots & 0 & \dots & 1 & \\
\hline
\multicolumn{6}{c}{0} & 0
\end{array}\right)
\end{array}, \qquad (5.16)$$

whose elements are zeros with the exception of those occupying the positions (1, 1), (2, 2), ..., (r, r), each of which is equal to unity.

Interchanging the columns in (5.15) so that the k_1th column replaces the first one, the k_2th column replaces the second one, etc., until the k_rth column replaces the rth column, we obtain a matrix of the form

$$\begin{pmatrix} \tilde{\alpha}_{11} & \tilde{\alpha}_{12} & \ldots & \tilde{\alpha}_{1r} & \ldots & \tilde{\alpha}_{1n} \\ & \tilde{\alpha}_{22} & \ldots & \tilde{\alpha}_{2r} & \ldots & \tilde{\alpha}_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & & \tilde{\alpha}_{rr} & \ldots & \tilde{\alpha}_{rn} \\ & 0 & & & & \end{pmatrix}, \tag{5.17}$$

where $\tilde{\alpha}_{11} \neq 0$, $\tilde{\alpha}_{22} \neq 0$, ..., $\tilde{\alpha}_{rr} \neq 0$.

For example, interchanging the third and fifth columns in $\mathbf{A}_5$, we obtain

$$\mathbf{A}_6 = \begin{pmatrix} 1 & -3 & 7 & 0 & 5 \\ 0 & 4 & 13 & 1 & 9 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Adding multiples of the first column of (5.17) by $-\dfrac{\tilde{\alpha}_{1j}}{\tilde{\alpha}_{11}}$ $(j = 2, 3, \ldots, n)$ to the corresponding columns, we obtain a matrix of the form

$$\begin{pmatrix} \tilde{\alpha}_{11} & 0 & \ldots & 0 & \ldots & 0 & \ldots & 0 \\ 0 & \tilde{\alpha}_{22} & \ldots & \tilde{\alpha}_{2j} & \ldots & \tilde{\alpha}_{2r} & \ldots & \tilde{\alpha}_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \ldots & 0 & \ldots & \tilde{\alpha}_{rr} & \ldots & \tilde{\alpha}_{rn} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \ldots & 0 & \ldots & 0 & \ldots & 0 \end{pmatrix}$$

where the first row contains only one nonzero element, $\tilde{\alpha}_{11}$.

Similarly, operating on the rows 2, 3, ..., r, we obtain the matrix

$$\begin{pmatrix} \tilde{\alpha}_{11} & & & & \\ & \tilde{\alpha}_{22} & & 0 & \\ & & \ddots & & \\ & & & \tilde{\alpha}_{ii} & \\ & & & & \ddots \\ 0 & & & & & \tilde{\alpha}_{rr} \end{pmatrix}. \tag{5.18}$$

By a sequence of elementary column operations of type (b) matrix (5.18) is reduced to the form (5.16).

It is easy to observe that the matrix $\mathbf{A}_6$ admits a representation of the form

$$\mathbf{A}_7 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

whence we obtain

$$\mathbf{A}_8 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Elementary matrices. The elementary operations which are summarized above are closely associated with square matrices called *elementary matrices.* These are of the following types:

(a) matrices obtained from the corresponding identity matrices by interchanging any two rows. For instance, the matrix

$$\mathbf{P}_{ij} = \begin{array}{cc} \begin{array}{ccccc} & i & & j & \end{array} & \\ \left(\begin{array}{ccccc} 1\ \ddots\ 1 & \vdots & & \vdots & \\ \cdots & 0 & \cdots & 1 & \cdots \\ & \vdots & 1\ \ddots\ 1 & \vdots & \\ \cdots & 1 & \cdots & 0 & \cdots \\ & \vdots & & \vdots & 1\ \ddots\ 1 \end{array}\right) & \begin{array}{c} \\ i \\ \\ j \\ \\ \end{array} \end{array}$$

is obtained from the identity matrix

$$\begin{array}{cc} \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & 0 & \\ & & 1 & & & & \\ & & & \ddots & & & \\ & & & & 1 & & \\ & 0 & & & & \ddots & \\ & & & & & & 1 \end{pmatrix} & \begin{array}{c} \\ \\ i \\ \\ j \\ \\ \\ \end{array} \end{array}$$

by interchanging the ith and jth rows. Observe that off-diagonal elements of $\mathbf{P}_{ij}$ are zeros with the exceptions of the elements in the (i, j)th and (j, i)th positions.

(b) Matrices obtained from the corresponding identity matrices by substituting a nonzero scalar for a diagonal element. For example, the matrix

$$\mathbf{D}_j = \begin{array}{cc} \begin{pmatrix} 1 & & & \vdots & & & \\ & \ddots & & \vdots & & & \\ & & 1 & \vdots & & & \\ \cdots & \cdots & \cdots & \beta & \cdots & \cdots & \cdots \\ & & & \vdots & 1 & & \\ & & & \vdots & & \ddots & \\ & & & \vdots & & & 1 \end{pmatrix} & \begin{array}{c} \\ \\ \\ j \\ \\ \\ \\ \end{array} \\ \begin{array}{ccccccc} & & & j & & & \end{array} & \end{array}$$

differs from the identity matrix by the element $\beta \neq 0$ in the (j, j)th position. Notice that all off-diagonal elements of $\mathbf{D}_j$ are zeros.

(c) Matrices that differ from the corresponding identity matrices by one off-diagonal element. For example, the matrix

$$\mathbf{L}_{ij} = \begin{array}{cc} \begin{array}{ccccccc} & & i & & & & \end{array} & \\ \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & \vdots & \ddots & & & \\ \cdots & \cdots & \gamma & \cdots & 1 & \cdots & \cdots \\ & & \vdots & & & \ddots & \\ & & \vdots & & & & 1 \end{pmatrix} & \begin{array}{c} \\ \\ \\ \\ j \\ \\ \\ \end{array} \end{array}$$

differs from the unit matrix by the element γ in the (j, i)th position, and the matrix

$$\mathbf{R}_{ij} = \begin{array}{cc} \begin{array}{ccccccc} & & & & j & & \end{array} & \\ \begin{pmatrix} 1 & & & & \vdots & & \\ & \ddots & & & \vdots & & \\ \cdots & \cdots & 1 & \cdots & \gamma & \cdots & \cdots \\ & & & \ddots & \vdots & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix} & \begin{array}{c} \\ \\ i \\ \\ \\ \\ \\ \end{array} \end{array}$$

also differs from the identity matrix by the same element but located in the (i, j)th position. Observe that off-diagonal elements of $\mathbf{L}_{ij}$ and $\mathbf{R}_{ij}$ are zeros with the exception of the element γ.

We point out here that for any matrix each elementary operation is equivalent to pre- or post-multiplication of the matrix by a suitable elementary matrix.

Theorem 5.2. *Elementary operations on a matrix are equivalent to pre- and post-multiplication of the matrix by elementary matrices.*

Let **A** be an arbitrary matrix and $\mathbf{P}_{ij}$, $\mathbf{D}_j$, $\mathbf{L}_{ij}$ and $\mathbf{R}_{ij}$ be elementary matrices given above. Then

(a) interchanging the *i*th and *j*th rows of **A** is equivalent to pre-multiplying **A** by $\mathbf{P}_{ij}$,

(b) multiplying one particular row of **A** by a nonzero scalar β is equivalent to pre-multiplication of **A** by $\mathbf{D}_j$,

(c) adding one row of **A** multiplied by a scalar γ to another row is equivalent to pre-multiplying **A** by $\mathbf{L}_{ij}$,

(a′) interchanging the *i*th and *j*th columns of **A** is equivalent to post-multiplication of **A** by $\mathbf{P}_{ij}$,

(b′) multiplying one particular column of **A** by a nonzero scalar β is equivalent to post-multiplication of **A** by $\mathbf{D}_j$,

(c′) adding one column of **A** multiplied by a scalar γ to another column is equivalent to post-multiplication of **A** by $\mathbf{R}_{ij}$.

◄ For simplicity we consider a square matrix of order 3

$$\mathbf{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix}.$$

Recall that pre-multiplying **A** by the matrix

$$\mathbf{P}_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

gives

$$\mathbf{B} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \end{pmatrix}$$

and post-multiplying **A** by $\mathbf{P}_{23}$ gives

$$\mathbf{C} = \begin{pmatrix} \alpha_{11} & \alpha_{13} & \alpha_{12} \\ \alpha_{21} & \alpha_{23} & \alpha_{22} \\ \alpha_{31} & \alpha_{33} & \alpha_{32} \end{pmatrix}.$$

It is easy to see that the matrix **B** differs from **A** by the order of rows and the matrix **C** differs from **A** by the order of columns.

Similarly we can verify the conditions (a) and (a′) for the matrices

$$\mathbf{P}_{12} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{P}_{13} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Let us multiply $\mathbf{A}$ by the matrix

$$\mathbf{D}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Pre-multiplying $\mathbf{A}$ by $\mathbf{D}_2$, we have

$$\mathbf{D}_2\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix}$$

$$= \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \beta\alpha_{21} & \beta\alpha_{22} & \beta\alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix}.$$

Post-multiplying $\mathbf{A}$ by $\mathbf{D}_2$, we obtain

$$\mathbf{A}\mathbf{D}_2 = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \alpha_{11} & \beta\alpha_{12} & \alpha_{13} \\ \alpha_{21} & \beta\alpha_{22} & \alpha_{23} \\ \alpha_{31} & \beta\alpha_{32} & \alpha_{33} \end{pmatrix}.$$

Similarly, we verify the conditions (b) and (b′) for the elementary matrices

$$\mathbf{D}_1 = \begin{pmatrix} \beta & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{D}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \beta \end{pmatrix}.$$

Notice that the conditions (c) and (c′) can easily be verified in a similar way. ►

5.2 Determinants

We associate with a square matrix a single number called a *determinant* according to the following rules:

The determinant of the matrix of order 1

$$(\alpha_{11})$$

is equal to α_{11}.

The determinant of the matrix of order 2

$$\begin{pmatrix}\alpha_{11} & \alpha_{12}\\ \alpha_{21} & \alpha_{22}\end{pmatrix}$$

is a number equal to $\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}$.

In symbols, we write

$$\det\begin{pmatrix}\alpha_{11} & \alpha_{12}\\ \alpha_{21} & \alpha_{22}\end{pmatrix} = \begin{vmatrix}\alpha_{11} & \alpha_{12}\\ \alpha_{21} & \alpha_{22}\end{vmatrix} = \alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}. \tag{5.19}$$

The determinant of the matrix of order 3

$$\begin{pmatrix}\alpha_{11} & \alpha_{12} & \alpha_{13}\\ \alpha_{21} & \alpha_{22} & \alpha_{23}\\ \alpha_{31} & \alpha_{32} & \alpha_{33}\end{pmatrix}$$

is a number equal to

$$\alpha_{11}\begin{vmatrix}\alpha_{22} & \alpha_{23}\\ \alpha_{32} & \alpha_{33}\end{vmatrix} - \alpha_{21}\begin{vmatrix}\alpha_{12} & \alpha_{13}\\ \alpha_{32} & \alpha_{33}\end{vmatrix} + \alpha_{31}\begin{vmatrix}\alpha_{12} & \alpha_{13}\\ \alpha_{22} & \alpha_{23}\end{vmatrix}.$$

Using (5.19), we obtain

$$\begin{aligned}\det\begin{pmatrix}\alpha_{11} & \alpha_{12} & \alpha_{13}\\ \alpha_{21} & \alpha_{22} & \alpha_{23}\\ \alpha_{31} & \alpha_{32} & \alpha_{33}\end{pmatrix} &= \begin{vmatrix}\alpha_{11} & \alpha_{12} & \alpha_{13}\\ \alpha_{21} & \alpha_{22} & \alpha_{23}\\ \alpha_{31} & \alpha_{32} & \alpha_{33}\end{vmatrix}\\ &= \alpha_{11}\alpha_{22}\alpha_{33} + \alpha_{21}\alpha_{32}\alpha_{13} + \alpha_{31}\alpha_{12}\alpha_{23}\\ &\quad - \alpha_{11}\alpha_{32}\alpha_{23} - \alpha_{21}\alpha_{12}\alpha_{33} - \alpha_{31}\alpha_{22}\alpha_{13}.\end{aligned} \tag{5.20}$$

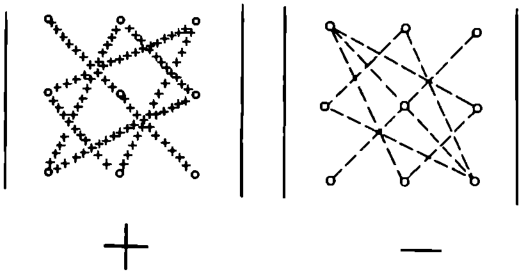


Fig. 5.4

Figure 5.4 is helpful to memorize formula (5.20). The drawing on the left refers to the arrangement of the positive terms in (5.20) and that on the right to the arrangement of the negative terms.

Suppose that we have defined determinants for matrices of orders less than *n*.

The *determinant of the matrix of order n*

$$\mathbf{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{n1} & \alpha_{n2} & \ldots & \alpha_{nn} \end{pmatrix} \tag{5.21}$$

is a number equal to

$$\mathbf{D} = \alpha_{11}M_{11} - \alpha_{21}M_{21} + \ldots + (-1)^{n+1}\alpha_{n1}M_{n1}, \tag{5.22}$$

where M_{i1} ($i = 1, 2, \ldots, n$) is the determinant of order $n - 1$

$$\begin{vmatrix} \alpha_{12} & \alpha_{13} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{23} & \ldots & \alpha_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{i-1,2} & \alpha_{i-1,3} & \ldots & \alpha_{i-1,n} \\ \alpha_{i+1,2} & \alpha_{i+1,3} & \ldots & \alpha_{i+1,n} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{n2} & \alpha_{n3} & \ldots & \alpha_{nn} \end{vmatrix} \tag{5.23}$$

obtained by deleting from $\mathbf{A}$ the ith row and the first column.

In symbols, we write

$$D = \det \mathbf{A} = |\mathbf{A}| = \begin{vmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{n1} & \alpha_{n2} & \ldots & \alpha_{nn} \end{vmatrix}. \tag{5.24}$$

Formula (5.22) defines the *expansion of determinant with respect to the first column.*

It is easy to see that for $n = 2$ and $n = 3$ formula (5.22) gives results identical to those given by (5.19) and (5.20), respectively.

For example, for $n = 3$ we have

$$\begin{aligned} D = {} & \alpha_{11}(\alpha_{22}\alpha_{33} - \alpha_{32}\alpha_{23}) - \alpha_{21}(\alpha_{12}\alpha_{33} - \alpha_{32}\alpha_{13}) \\ & + \alpha_{31}(\alpha_{12}\alpha_{23} - \alpha_{22}\alpha_{13}). \end{aligned} \tag{5.25}$$

We abbreviate (5.22) by writing

$$D = \sum_{i=1}^{n} (-1)^{i+1}\alpha_{i1}M_{i1}. \tag{5.26}$$

Example. Compute the determinant of the *triangular* matrix

$$\mathbf{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \ldots & \alpha_{1n} \\ 0 & \alpha_{22} & \alpha_{23} & \ldots & \alpha_{2n} \\ 0 & 0 & \alpha_{33} & \ldots & \alpha_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \ldots & \alpha_{nn} \end{pmatrix}.$$

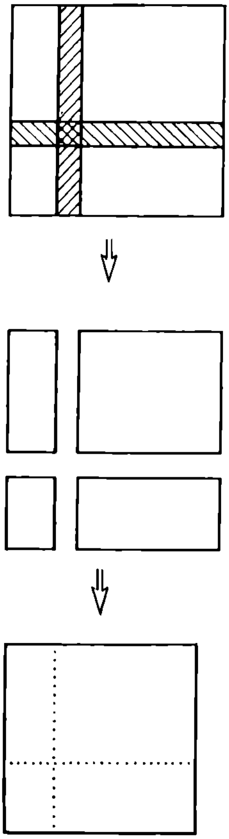
◄ We have

$$|\mathbf{A}| = \alpha_{11} \begin{vmatrix} \alpha_{22} & \alpha_{23} & \dots & \alpha_{2n} \\ 0 & \alpha_{33} & \dots & \alpha_{3n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \alpha_{nn} \end{vmatrix} = \alpha_{11}\alpha_{22} \begin{vmatrix} \alpha_{33} & \dots & \alpha_{3n} \\ \dots & \dots & \dots \\ 0 & \dots & \alpha_{nn} \end{vmatrix}$$

$$= \dots = \alpha_{11}\alpha_{22} \dots \alpha_{nn}.$$

Whence we infer that the ***determinant of the triangular matrix is equal to the product of the elements on the principal diagonal.*** ►

The determinant of a matrix of order $n - 1$ obtained from the matrix **A** by deleting the ith row and the jth column is called the ***minor*** M_{ij} of **A** (Fig. 5.5). For instance, the minor M_{i1} is the determinant of the matrix obtained from **A** by deleting the ith row and the first column.



Minor M_{ij} obtained from matrix A by deleting the ith row and the jth column

Fig. 5.5

Consider n numbers

$$D_j = \sum_{i=1}^{n} (-1)^{i+j}\alpha_{ij}M_{ij} \quad (j = 1, 2, \ldots, n). \tag{5.27}$$

It is easy to see that for $j = 1$ formula (5.27) coincides with (5.26). Let us prove that $D = D_1 = D_2 = \ldots = D_n$.

◄ We confine ourselves to the case when $n = 3$. By setting $j = 2$ in (5.27), we have

$$D_2 = -\alpha_{12}M_{12} + \alpha_{22}M_{22} - \alpha_{32}M_{32}, \tag{5.28}$$

where

$$M_{12} = \begin{vmatrix} \alpha_{21} & \alpha_{23} \\ \alpha_{31} & \alpha_{33} \end{vmatrix} = \alpha_{21}\alpha_{33} - \alpha_{31}\alpha_{23},$$

$$M_{22} = \begin{vmatrix} \alpha_{11} & \alpha_{13} \\ \alpha_{31} & \alpha_{33} \end{vmatrix} = \alpha_{11}\alpha_{33} - \alpha_{31}\alpha_{13}, \tag{5.29}$$

$$M_{32} = \begin{vmatrix} \alpha_{11} & \alpha_{13} \\ \alpha_{21} & \alpha_{23} \end{vmatrix} = \alpha_{11}\alpha_{23} - \alpha_{13}\alpha_{21}.$$

Whence we obtain

$$\begin{aligned} D_2 = {} & -\alpha_{12}(\alpha_{12}\alpha_{33} - \alpha_{31}\alpha_{23}) + \alpha_{22}(\alpha_{11}\alpha_{33} - \alpha_{13}\alpha_{31}) \\ & - \alpha_{32}(\alpha_{11}\alpha_{23} - \alpha_{13}\alpha_{21}). \end{aligned} \tag{5.30}$$

Comparing (5.25) and (5.30), we conclude that $D = D_2$. The identity $D = D_3$ is verified in a similar way. ►

Remark. In general the identities $D = D_1 = D_2 = \ldots = D_n$ are verified by computing determinants of orders $n - 1$, $n - 2$, etc.

Therefore we have proved that the *expansion of determinant with respect to the jth column* is given by

$$D = \sum_{i=1}^{n} (-1)^{i+j}\alpha_{ij}M_{ij} \quad (j = 1, 2, \ldots, n). \tag{5.31}$$

We call the number:

$$A_{ij} = (-1)^{i+j}M_{ij} \tag{5.32}$$

the *cofactor* of the element α_{ij} in $|\mathbf{A}|$.

Notice that the cofactor A_{ij} of α_{ij} is independent of the value of α_{ij}. In other words, A_{ij} remains unchanged when the element α_{ij} is replaced by β_{ij} in $\mathbf{A}$.

Substituting (5.32) into (5.31), we may write

$$D = \sum_{i=1}^{n} \alpha_{ij}A_{ij} = \alpha_{1j}A_{1j} + \alpha_{2j}A_{2j} + \ldots + \alpha_{nj}A_{nj}$$

$$(j = 1, 2, \ldots, n). \tag{5.33}$$

Whence we conclude that the *determinant of a square matrix is equal to the sum of the elements of any column multiplied by their respective cofactors.*

Similarly to (5.27) we may introduce the numbers

$$\Delta_i = \sum_{j=1}^{n} (-1)^{i+j}\alpha_{ij}M_{ij} \quad (i = 1, 2, \ldots, n) \tag{5.34}$$

such that $\Delta_1 = \Delta_2 = \ldots = \Delta_n$. To verify these identities it suffices to interchange the rows and columns in **A**, i.e., to transpose **A**, and apply to $\mathbf{A}'$ the same reasoning as above. In conclusion we shall prove the following theorem.

Theorem 5.3. *For any square matrix of order n there holds*

$$D = \sum_{j=1}^{n} (-1)^{i+j}\alpha_{ij}M_{ij} \quad (i = 1, 2, \ldots, n), \tag{5.35}$$

i.e., the determinant of any square matrix of order n admits expansion with respect to the ith row.

◄ It is sufficient to prove that

$$\Delta_1 = D. \tag{5.36}$$

Again we confine ourselves to $n = 3$.

Using (5.34), we have

$$\Delta_1 = \alpha_{11}M_{11} - \alpha_{12}M_{12} + \alpha_{13}M_{13} = \alpha_{11}(\alpha_{22}\alpha_{33} - \alpha_{23}\alpha_{32}) - \alpha_{12}(\alpha_{21}\alpha_{33} - \alpha_{23}\alpha_{31}) + \alpha_{13}(\alpha_{21}\alpha_{32} - \alpha_{22}\alpha_{31}).$$

Comparing this expression with (5.25), we infer that $\Delta_1 = D$. ►

Remark. In general the identity $\Delta_1 = D$ is verified by computing determinants of orders $n - 1$, $n - 2$, etc.

Substituting (5.32) into (5.35), we arrive at

$$D = \sum_{j=1}^{n} \alpha_{ij}A_{ij} = \alpha_{i1}A_{i1} + \alpha_{i2}A_{i2} + \ldots + \alpha_{in}A_{in} \quad (i = 1, 2, \ldots, n). \tag{5.37}$$

Therefore, the *determinant of a square matrix of order n is equal to the sum of the elements of any row multiplied by their respective cofactors.*

Example. Compute the determinants of the elementary matrices $\mathbf{P}_{ij}$, $\mathbf{D}_j$ and $\mathbf{L}_{ij}$.

◄ The determinant of $\mathbf{P}_{ij}$ is

$$|\mathbf{P}_{ij}| = \begin{array}{|ccccccccccc|c}
 & & & i & & & & j & & & & \\
1 & & & \cdot & & & & & & & & \\
 & \ddots & & \vdots & & & & & & & & \\
 & & 1 & \cdot & & & & & & & & \\
 & & & 0 & & \cdots & & 1 & & & & i \\
 & & & & 1 & & & & & & & \\
 & & & & & \ddots & & & & & & \cdot \\
 & & & & & & 1 & & & & & \\
 & & & 1 & & & & 0 & \cdots & & & j \\
 & & & & & & & & 1 & & & \\
 & & & & & & & & & \ddots & & \\
 & & & & & & & & & & 1 &
\end{array}$$

Expanding $\mathbf{P}_{ij}$ with respect to the first row, we have a determinant of order $n - 1$ equal to $|\mathbf{P}_{ij}|$. Then expanding the determinant of order $n - 1$ with respect to its first row, we obtain a determinant of order $n - 2$ which is also equal to $|\mathbf{P}_{ij}|$. It is easy to see that going on with this procedure we obtain

$$|\mathbf{P}_{ij}| = \begin{vmatrix}
0 & \cdots & & & 1 & & & \\
\cdot & 1 & & & & & & \\
\vdots & & \ddots & & & & & \\
 & & & 1 & & & & \\
1 & \cdot & & \cdot & 0 & & & \\
 & & & & & 1 & & \\
 & & & & & & \ddots & \\
 & & & & & & & 1
\end{vmatrix}$$

after $i - 1$ steps.

Using this procedure further, we finally arrive at

$$|\mathbf{P}_{ij}| = \begin{vmatrix} 0 & 1 \\ 1 & 0 \end{vmatrix} = -1.$$

Recall that $\mathbf{D}_j$ is a diagonal matrix of order n such that $\alpha_{jj} = \beta$ and the other diagonal elements are unities. Then

$$|\mathbf{D}_j| = \beta.$$

For the matrix $\mathbf{L}_{ij}$ we have

$$|\mathbf{L}_{ij}| = 1. \blacktriangleright$$

Properties of determinants. Chapter 1 represents a detailed account of properties of the second- and third-order determinants. Here we point out properties possessed by an arbitrary determinant of order n.

Linear property. Suppose that we are given a determinant D such that one of its rows is a linear combination of the other two rows. Let

$$D = \begin{vmatrix} \alpha_{11} & \alpha_{12} & & \alpha_{1n} \\ \lambda\beta_1 + \mu\gamma_1 & \lambda\beta_2 + \mu\gamma_2 & \dots & \lambda\beta_n + \mu\gamma_n \\ \dots & \dots & \dots & \dots \\ \alpha_{n1} & \alpha_{n2} & & \alpha_{nn} \end{vmatrix} \; i.$$

Then

$$D = \lambda D' + \mu D'',$$

where the determinants

$$D' = \begin{vmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \dots & \dots & \dots & \dots \\ \beta_1 & \beta_2 & \dots & \beta_n \\ \dots & \dots & \dots & \dots \\ \alpha_{n1} & \alpha_{n2} & & \alpha_{nn} \end{vmatrix} \; i \quad \text{and} \quad D'' = \begin{vmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \dots & \dots & \dots & \dots \\ \gamma_1 & \gamma_2 & \dots & \gamma_n \\ \dots & \dots & \dots & \dots \\ \alpha_{n1} & \alpha_{n2} & & \alpha_{nn} \end{vmatrix} \; i$$

differ from D by their ith rows.

◀ To verify this property it suffices to expand D, D' and D'' with respect to their ith rows. Observe that the cofactors of $\lambda\beta_j + \mu\gamma_j$, β_j and γ_j are all equal to A_{ij}. Then using (5.37) we have

$$D = \sum_{j=1}^{n} (\lambda\beta_j + \mu\gamma_j)A_{ij},$$

$$D' = \sum_{j=1}^{n} \beta_j A_{ij} \quad \text{and} \quad D'' = \sum_{j=1}^{n} \gamma_j A_{ij}.$$

Whence

$$D = \lambda D' + \mu D''. \blacktriangleright$$

Property of antisymmetry. If a determinant $\tilde{D}$ is obtained from a determinant D by interchanging any two rows in D, then

$$\tilde{D} = -D.$$

◄ Suppose that

$$\tilde{D} = \begin{pmatrix} \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \dots & \dots & \dots & \dots \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} \end{pmatrix}$$

is obtained from

$$D = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \dots & \dots & \dots & \dots \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} \end{pmatrix}$$

by interchanging the first and second rows in D.

Expand D with respect to the second row and $\tilde{D}$ with respect to the first row. Using (5.35), we obtain

$$D = -\alpha_{21}M_{21} + \alpha_{22}M_{22} + \dots + (-1)^{2+n}\alpha_{2n}M_{2n}$$

and

$$\tilde{D} = \alpha_{21}M_{21} - \alpha_{22}M_{22} + \dots + (-1)^{1+n}\alpha_{2n}M_{2n}.$$

Whence

$$\tilde{D} = -D.$$ ►

Similarly we may prove this property in the case when $\tilde{D}$ is obtained from D by interchanging any two rows.

Transposition of a determinant. The operation of transposing a matrix **A** leaves the determinant of **A** unchanged, i.e.

$$|\mathbf{A}'| = |\mathbf{A}|.$$

◄ Expanding $|\mathbf{A}|$ with respect to the first row and $|\mathbf{A}'|$ with respect to the first column, we infer that $|\mathbf{A}'| = |\mathbf{A}|$. ►

Notice that this property implies the equivalence of rows and columns of any determinant. So the above two properties may be described in terms of columns of determinants.

Multiplication of determinants. Let **A** and **B** be two square matrices of the same order. Then

$$|\mathbf{AB}| = |\mathbf{A}|\cdot|\mathbf{B}|,$$

that is, the determinant of the product of two square matrices of the same order is equal to the product of the respective determinants.

Now we shall mention some properties of determinants which are helpful in computing determinants.

(a) The determinant with two identical rows is equal to zero.

◄ Indeed, by property of antisymmetry interchanging any two rows reverses the sign of determinant. However, interchanging two identical rows leaves the determinant unchanged. Hence $D = -D$. This means that $D = 0$. ►

(b) Multiplying a row by a scalar is equivalent to multiplying the determinant by the same scalar.

◄ This follows from linear property provided that $\mu = 0$. ►

(c) The determinant with a zero row is equal to zero.

◄ To prove this property it suffices to expand the determinant with respect to the zero row. ►

(d) The determinant with two proportional rows is equal to zero.

◄ By property (b) the common factor may be taken outside the determinant. Then two rows become identical. Hence the determinant is equal to zero. ►

(e) Adding one row multiplied by a scalar to another row leaves the determinant unchanged.

◄ By linear property this operation results in the determinant equal to the sum of two determinants, i.e., the original determinant and that with two proportional rows. By property (d) the latter is equal to zero. Therefore the determinant remains unchanged. ►

In general a *determinant remains unchanged when a linear combination of rows is added to any of its rows.* The same is true for the columns of a determinant.

Example. Prove that the sum of elements of one row multiplied by cofactors of the respective elements of another row is equal to zero.

◄ Replacing the kth row by the ith row in

$$D = \begin{array}{|cccc|c}
\alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} & \\
\cdots & \cdots & \cdots & \cdots & \\
\alpha_{i1} & \alpha_{i2} & \dots & \alpha_{in} & i \\
\cdots & \cdots & \cdots & \cdots & \\
\alpha_{k1} & \alpha_{k2} & \dots & \alpha_{kn} & k \\
\cdots & \cdots & \cdots & \cdots & \\
\alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} &
\end{array} \qquad (i \neq k)$$

we obtain

$$\tilde{D} = \begin{array}{|cccc|c}
\alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} & \\
\cdots & \cdots & \cdots & \cdots & \\
\alpha_{i1} & \alpha_{i2} & \dots & \alpha_{in} & i \\
\cdots & \cdots & \cdots & \cdots & \\
\alpha_{i1} & \alpha_{i2} & \dots & \alpha_{in} & k \\
\cdots & \cdots & \cdots & \cdots & \\
\alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} &
\end{array}$$

Observe that the ith and kth rows in $\tilde{D}$ are identical. By property (a), $\tilde{D} = 0$. Then expanding $\tilde{D}$ with respect to the kth row, we arrive at the desired identity

$$\alpha_{i1}A_{k1} + \alpha_{i2}A_{k2} + \ldots + \alpha_{in}A_{kn} = 0 \quad (i \neq k).$$

(Recall that cofactors of elements in any row do not depend on the values of these elements.) ►

Computing determinants. When computing determinants we widely use elementary (row and column) operations. It is important to point out that being subjected to elementary operations either (a) or (b) or (c) the determinant either reverses its sign or is multiplied by a scalar or remains unchanged, respectively. Therefore the elementary operations are closely related to the properties of determinants.

Example. Compute the determinant of

$$\mathbf{A} = \begin{pmatrix} 2 & -5 & 4 & 3 \\ 3 & -4 & 7 & 4 \\ 4 & -10 & 8 & 3 \\ -3 & 2 & -5 & 3 \end{pmatrix}.$$

◄ Observe that $\alpha_{11} \neq 0$ and divides only two elements of the first column with non-zero remainders. To avoid division of elements of the matrix we multiply the second row by -2, the third row by -1 and the fourth row by 2. Then we obtain

$$(-2)(-1)2\,|\mathbf{A}| = \begin{vmatrix} 2 & -5 & 4 & 3 \\ -6 & 8 & -14 & -8 \\ -4 & 10 & -8 & -3 \\ -6 & 4 & -10 & 6 \end{vmatrix}.$$

Step 1. Adding multiples by 3, 2 and 3 of the first row to the second, third and fourth rows, respectively, we obtain

$$4\,|\mathbf{A}| = \begin{vmatrix} 2 & -5 & 4 & 3 \\ 0 & -7 & -2 & 1 \\ 0 & 0 & 0 & 3 \\ 0 & -11 & 2 & 15 \end{vmatrix}.$$

Step 2. To avoid division we multiply the fourth row by -7. Then we have

$$(-7)4\,|\mathbf{A}| = \begin{vmatrix} 2 & -5 & 4 & 3 \\ 0 & -7 & -2 & 1 \\ 0 & 0 & 0 & 3 \\ 0 & 77 & -14 & -105 \end{vmatrix}.$$

Adding the second row multiplied by 11 to the fourth row gives

$$-28\,|\mathbf{A}| = \begin{vmatrix} 2 & -5 & 4 & 3 \\ 0 & -7 & -2 & 1 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & -36 & -94 \end{vmatrix}.$$

Step 3. Interchanging the third and fourth rows gives

$$28\,|\mathbf{A}| = \begin{vmatrix} 2 & -5 & 4 & 3 \\ 0 & -7 & -2 & 1 \\ 0 & 0 & -36 & -94 \\ 0 & 0 & 0 & 3 \end{vmatrix}.$$

Computing the determinant of the triangular matrix given above, we obtain

$$28\,|\mathbf{A}| = 2(-7)(-36)3$$

and finally

$$|\mathbf{A}| = 54.$$ ►

5.3 Inverse Matrices

A square matrix is called *nonsingular* if its determinant is not equal to zero.

Let $\mathbf{A} = (\alpha_{ij})$ be a nonsingular matrix of order n and let $\mathbf{B}$ be a square matrix of order n such that the element of $\mathbf{B}$ in the position where the ith row and jth column meet is equal to the cofactor A_{ji} of α_{ji} in $\mathbf{A}$, i.e., $\mathbf{B} = (\beta_{ij})$ where $\beta_{ij} = A_{ji}$. Therefore the matrices $\mathbf{A}$ and $\mathbf{B}$ are related as shown below:

$$\mathbf{A} = \begin{array}{cc} \begin{pmatrix} \alpha_{11} & \dots & \alpha_{1i} & \dots & \alpha_{1n} \\ \alpha_{21} & \dots & \alpha_{2i} & \dots & \alpha_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha_{j1} & \dots & \alpha_{ji} & \dots & \alpha_{jn} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha_{n1} & \dots & \alpha_{ni} & \dots & \alpha_{nn} \end{pmatrix} & \begin{array}{c} \\ \\ \\ j; \\ \\ \\ \end{array} \\ \begin{array}{ccccc} \phantom{\alpha_{11}} & & i & & \phantom{\alpha_{1n}} \end{array} & \end{array}$$

$$\mathbf{B} = \begin{array}{cc} \begin{pmatrix} A_{11} & \dots & A_{j1} & \dots & A_{n1} \\ A_{12} & \dots & A_{j2} & \dots & A_{n2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{1i} & \dots & A_{ji} & \dots & A_{ni} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{1n} & \dots & A_{jn} & \dots & a_{nn} \end{pmatrix} & \begin{array}{c} \\ \\ \\ i. \\ \\ \\ \end{array} \\ \begin{array}{ccccc} \phantom{A_{11}} & & j & & \phantom{A_{n1}} \end{array} & \end{array}$$

The matrix **B** possesses the following property

$$\mathbf{AB} = \mathbf{BA} = \begin{pmatrix} |\mathbf{A}| & & 0 \\ & \ddots & \\ 0 & & |\mathbf{A}| \end{pmatrix} = |\mathbf{A}| \cdot \mathbf{I}. \qquad (5.38)$$

By way of illustration we show that

$$\mathbf{AB} = |\mathbf{A}| \cdot \mathbf{I}.$$

◄ The element of **AB** occupying the position where the *i*th row and the *j*th column meet is given by

$$\gamma_{ij} = \sum_{k=1}^{n} \alpha_{ik} A_{jk}.$$

Setting $i = j$, we see that γ_{ii} is the result of the expansion of the determinant of **A** with respect to the *i*th row, i.e.,

$$\gamma_{ii} = \sum_{k=1}^{n} \alpha_{ik} A_{ik} = |\mathbf{A}|.$$

In other words, each diagonal element of **AB** is equal to the value of the determinant of **A**.

By the exercise considered in the previous section, we have

$$\gamma_{ij} = 0$$

for each off-diagonal element in **AB**. ►

Similar reasoning is applicable to the identity $\mathbf{BA} = |\mathbf{A}| \cdot \mathbf{I}$.

The matrix

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}\mathbf{B} = \begin{pmatrix} \frac{A_{11}}{|\mathbf{A}|} & \frac{A_{j1}}{|\mathbf{A}|} & \frac{A_{n1}}{|\mathbf{A}|} \\ \cdots & \cdots & \cdots \\ \frac{A_{1i}}{|\mathbf{A}|} & \frac{A_{ji}}{|\mathbf{A}|} & \frac{A_{ni}}{|\mathbf{A}|} \\ \cdots & \cdots & \cdots \\ \frac{A_{1n}}{|\mathbf{A}|} & \frac{A_{jn}}{|\mathbf{A}|} & \frac{A_{nn}}{|\mathbf{A}|} \end{pmatrix} \qquad (5.39)$$

is called the *inverse of the matrix* **A**.

From (5.38) it follows that

$$\mathbf{AA}^{-1} = \mathbf{I}, \quad \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}. \qquad (5.40)$$

This means that the matrix $\mathbf{A}^{-1}$ may be considered as a solution satisfying two matrix equations simultaneously

$$\mathbf{AX} = \mathbf{I} \quad \text{and} \quad \mathbf{XA} = \mathbf{I},$$

where

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{in} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nj} & \dots & x_{nn} \end{pmatrix}$$

is an unknown matrix.

Let us show that $\mathbf{A}^{-1}$ is the only solution satisfying both these equations simultaneously.

Assume that there exists a matrix $\mathbf{C}$ such that

$$\mathbf{AC} = \mathbf{I} \quad \text{and} \quad \mathbf{CA} = \mathbf{I}.$$

Pre-multiplying the former identity by $\mathbf{A}^{-1}$ and post-multiplying the latter by $\mathbf{A}^{-1}$, we have

$$\mathbf{A}^{-1}(\mathbf{AC}) = (\mathbf{A}^{-1})\mathbf{I} \quad \text{and} \quad (\mathbf{CA})\mathbf{A}^{-1} = \mathbf{I}(\mathbf{A}^{-1}). \tag{5.41}$$

By the associative law of multiplication we may write

$$\mathbf{A}^{-1}(\mathbf{AC}) = (\mathbf{A}^{-1}\mathbf{A})\mathbf{C} \quad \text{and} \quad (\mathbf{CA})\mathbf{A}^{-1} = \mathbf{C}(\mathbf{AA}^{-1}).$$

Substituting these identities into (5.41) and using (5.40), we obtain

$$\mathbf{IC} = \mathbf{A}^{-1}\mathbf{I} \quad \text{and} \quad \mathbf{CI} = \mathbf{IA}^{-1}.$$

Whence, (5.6) gives

$$\mathbf{C} = \mathbf{A}^{-1}.$$

Computing inverse matrices. We shall mention here an effective and rather straightforward method for computing an inverse matrix by successive use of elementary (row and column) operations. We begin with theoretical background of the method, known as *pivotal condensation.*

Theorem 5.4. *By a sequence of elementary row (column) operations an arbitrary nonsingular matrix is reducible to the identity matrix.*

◄ By Theorem 5.1 any matrix is reducible by elementary row operations to a matrix of schematic form.

Suppose that we are given a square nonsingular matrix.

Let us show that this matrix may be reduced to the triangular matrix

$$\begin{pmatrix} \alpha_{11}^{(1)} & \alpha_{12}^{(1)} & \alpha_{13}^{(1)} & & \alpha_{1n}^{(1)} \\ 0 & \alpha_{22}^{(2)} & \alpha_{23}^{(2)} & & \alpha_{2n}^{(2)} \\ 0 & 0 & \alpha_{33}^{(3)} & \dots & \alpha_{3n}^{(3)} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \alpha_{nn}^{(n)} \end{pmatrix}, \tag{5.42}$$

where $\alpha_{11}^{(1)} \neq 0$, $\alpha_{22}^{(2)} \neq 0$, $\alpha_{33}^{(3)} \neq 0$, $\dots$, $\alpha_{nn}^{(n)} \neq 0$ (see formula (5.15)).

By Theorem 5.2 each elementary row operation is equivalent to premultiplication of the given matrix by a suitable elementary matrix. Recall that the determinants of the elementary matrices are not equal to zero. We have also seen that the determinant of the product of two square matrices is equal to the product of the determinants of these matrices. Thus multiplying a nonsingular matrix by any elementary matrix and, consequently, any elementary row operation, gives rise to a nonsingular matrix. This means that if the matrix (5.42) is obtained from the given nonsingular matrix by applying a sequence of elementary row operations, (5.42) is nonsingular and, consequently, its determinant

$$\alpha_{11}^{(1)}\cdot\alpha_{22}^{(2)}\cdot\alpha_{33}^{(3)}\cdot\ldots\cdot\alpha_{nn}^{(n)}$$

is distinct from zero. Whence we infer that

$$\alpha_{11}^{(1)}\neq 0,\quad \alpha_{22}^{(2)}\neq 0,\quad \alpha_{33}^{(3)}\neq 0,\ \ldots,\ \alpha_{nn}^{(n)}\neq 0.$$

If (5.42) is subjected to a sequence of elementary row operations of type (b), it becomes

$$\begin{pmatrix} 1 & \tilde{\alpha}_{12} & \tilde{\alpha}_{13} & \ldots & \tilde{\alpha}_{1n} \\ 0 & 1 & \tilde{\alpha}_{23} & \ldots & \tilde{\alpha}_{2n} \\ 0 & 0 & 1 & \ldots & \tilde{\alpha}_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix}. \tag{5.43}$$

Adding multiplies of the nth row of (5.43) by $-\tilde{\alpha}_{1n}$, $-\tilde{\alpha}_{2n}$, ..., $-\tilde{\alpha}_{n-1,n}$ to the 1st, 2nd, ..., $(n-1)$th rows, respectively, we obtain the matrix

$$\begin{pmatrix} 1 & \tilde{\alpha}_{12} & \tilde{\alpha}_{13} & \ldots & \tilde{\alpha}_{1,n-1} & 0 \\ 0 & 1 & \tilde{\alpha}_{23} & \ldots & \tilde{\alpha}_{2,n-1} & 0 \\ 0 & 0 & 1 & \ldots & \tilde{\alpha}_{3,n-1} & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix},$$

such that in its nth column all the elements but the last one, which is equal to 1, are zeros.

Similarly, adding multiplies of the $(n-1)$th row of the obtained matrix by $-\tilde{\alpha}_{1,n-1}$, $-\tilde{\alpha}_{2,n-1}$, ..., $-\tilde{\alpha}_{n-2,n-1}$ to the first $n-2$ rows, we obtain the matrix such that its $(n-1)$th and nth columns contain unities in the principal diagonal and zeros in the other rows. In the end of this process we add the multiple by $-\tilde{\alpha}_{12}$ of the second row of the matrix obtained after $n-1$ steps to the first row and finally arrive at the identity matrix

$$\begin{pmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{pmatrix}$$

and thus complete the proof. ►

In matrix form this theorem states the following.

Theorem 5.5. *Let* **A** *be an arbitrary nonsingular matrix. Then there exist elementary matrices* $\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_k$ *such that*

$$\mathbf{Q}_k \cdot \mathbf{Q}_{k-1} \cdot \ldots \cdot \mathbf{Q}_2 \cdot \mathbf{Q}_1 \cdot \mathbf{A} = \mathbf{I}. \tag{5.44}$$

Post-multiplying (5.44) by $\mathbf{A}^{-1}$, we obtain

$$\mathbf{Q}_k \cdot \mathbf{Q}_{k-1} \cdot \ldots \cdot \mathbf{Q}_2 \cdot \mathbf{Q}_1 = \mathbf{A}^{-1}.$$

Pivotal condensation. Suppose that we are given a nonsingular matrix **A** of order n. Let

$$(\mathbf{A} \mid \mathbf{I}) \tag{5.45}$$

be a matrix of order $n \times 2n$. If (5.45) is subjected to a sequence of elementary row operations or, equivalently, is premultiplied by a sequence of elementary matrices, the matrix **A** reduces to the identity matrix and the identity matrix **I** reduces to the matrix $\mathbf{A}^{-1}$ inverse to **A**. Thus, by elementary row operations (5.45) is reduced to

$$(\mathbf{I} \mid \mathbf{A}^{-1}).$$

It is easy to understand how we should proceed to compute the inverse of an arbitrary nonsingular square matrix $\mathbf{A} = (\alpha_{ij})$ $(i, j = 1, 2, \ldots, n)$:

1. Set up the matrix

$$(\mathbf{A} \mid \mathbf{I}) = \left(\begin{array}{cccc|cccc} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} & 1 & 0 & \ldots & 0 \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} & 0 & 1 & \ldots & 0 \\ \multicolumn{4}{c|}{\dotfill} & \multicolumn{4}{c}{\dotfill} \\ \alpha_{n1} & \alpha_{n2} & \ldots & \alpha_{nn} & 0 & 0 & \ldots & 1 \end{array}\right)$$

of order $n \times 2n$.

2. By elementary row operations reduce $(\mathbf{A} \mid \mathbf{I})$ to the form

$$(\tilde{\mathbf{A}} \mid \mathbf{B}) = \left(\begin{array}{cccc|cccc} 1 & \tilde{\alpha}_{12} & \ldots & \tilde{\alpha}_{1n} & \beta_{11} & \beta_{12} & \ldots & \beta_{1n} \\ 0 & 1 & \ldots & \tilde{\alpha}_{2n} & \beta_{21} & \beta_{22} & \ldots & \beta_{2n} \\ \multicolumn{4}{c|}{\dotfill} & \multicolumn{4}{c}{\dotfill} \\ 0 & 0 & \ldots & 1 & \beta_{n1} & \beta_{n2} & \ldots & \beta_{nn} \end{array}\right),$$

where $\tilde{\mathbf{A}}$ is the triangular matrix of order n obtained from **A**. (See the procedure used to prove Theorem 5.1.)

3. By elementary row operations reduce $(\tilde{\mathbf{A}} \,|\, \mathbf{B})$ to the form

$$(\mathbf{I} \,|\, \mathbf{C}) = \left(\begin{array}{cccc|cccc} 1 & 0 & \dots & 0 & \gamma_{11} & \gamma_{12} & \dots & \gamma_{1n} \\ 0 & 1 & \dots & 0 & \gamma_{21} & \gamma_{22} & \dots & \gamma_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & \gamma_{n1} & \gamma_{n2} & \dots & \gamma_{nn} \end{array}\right),$$

where $\mathbf{I}$ is the identity matrix obtained from $\tilde{\mathbf{A}}$. (See the procedure employed in Theorem 5.4.)

4. The matrix $\mathbf{C} = (\gamma_{ij})$ of order n $(i, j = 1, 2, \dots, n)$ in $(\mathbf{I} \,|\, \mathbf{C})$ is the inverse of the matrix $\mathbf{A}$, i.e.,

$$\mathbf{C} = \mathbf{A}^{-1}.$$

Example. Compute the matrix inverse to

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

◄ Set up the matrix of order 4 × 8

$$\mathbf{B} = \left(\begin{array}{cccc|cccc} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & -1 & -1 & 0 & 1 & 0 & 0 \\ 1 & -1 & 1 & -1 & 0 & 0 & 1 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 & 0 & 1 \end{array}\right).$$

Step 1. Subtract the first row from the other three rows:

$$\mathbf{B} \Rightarrow \mathbf{B}_1 = \left(\begin{array}{cccc|cccc} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -2 & -2 & -1 & 1 & 0 & 0 \\ 0 & -2 & 0 & -2 & -1 & 0 & 1 & 0 \\ 0 & -2 & -2 & 0 & -1 & 0 & 0 & 1 \end{array}\right).$$

Step 2. Observe that $\alpha_{22} = 0$. Interchange the second and third rows and subtract the second row thus obtained from the fourth row:

$$\mathbf{B}_1 \Rightarrow \mathbf{B}_2 = \left(\begin{array}{cccc|cccc} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & -2 & 0 & -2 & -1 & 0 & 1 & 0 \\ 0 & 0 & -2 & -2 & -1 & 1 & 0 & 0 \\ 0 & 0 & -2 & 2 & 0 & 0 & -1 & 1 \end{array}\right).$$

Step 3. Subtract the third row from the fourth one and divide each row by its diagonal element:

$$\mathbf{B}_2 \Rightarrow \mathbf{B}_3 = \left(\begin{array}{cccc|cccc} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1/2 & 0 & -1/2 & 0 \\ 0 & 0 & 1 & 1 & 1/2 & -1/2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1/4 & 1/4 & -1/4 & 1/4 \end{array}\right).$$

Step 4. Subtract the fourth row from the other three rows:

$$\mathbf{B}_3 \Rightarrow \mathbf{B}_4 = \left(\begin{array}{cccc|cccc} 1 & 1 & 1 & 0 & 3/4 & 1/4 & 1/4 & -1/4 \\ 0 & 1 & 0 & 0 & 1/4 & 1/4 & -1/4 & -1/4 \\ 0 & 0 & 1 & 0 & 1/4 & -1/4 & 1/4 & -1/4 \\ 0 & 0 & 0 & 1 & 1/4 & 1/4 & -1/4 & 1/4 \end{array}\right).$$

Step 5. Subtract the third row from the first one:

$$\mathbf{B}_4 \Rightarrow \mathbf{B}_5 = \left(\begin{array}{cccc|cccc} 1 & 1 & 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1/4 & 1/4 & -1/4 & -1/4 \\ 0 & 0 & 1 & 0 & 1/4 & -1/4 & 1/4 & -1/4 \\ 0 & 0 & 0 & 1 & 1/4 & 1/4 & -1/4 & 1/4 \end{array}\right).$$

Step 6. Subtract the second row from the first one:

$$\mathbf{B}_5 \Rightarrow \mathbf{B}_6 = \left(\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 1 & 0 & 0 & 1/4 & 1/4 & -1/4 & -1/4 \\ 0 & 0 & 1 & 0 & 1/4 & -1/4 & 1/4 & -1/4 \\ 0 & 0 & 0 & 1 & 1/4 & -1/4 & -1/4 & 1/4 \end{array}\right).$$

Whence we conclude that

$$\mathbf{A}^{-1} = \frac{1}{4}\mathbf{A}. \blacktriangleright$$

5.4 Rank of a Matrix

Suppose that we are given the matrix

$$\mathbf{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \dots & \dots & \dots & \dots \\ \alpha_{m1} & \alpha_{m2} & \dots & \alpha_{mn} \end{pmatrix}.$$

Let us choose in **A** k rows and k columns such that $i_1 < i_2 < \dots < i_k$ and $j_1 < j_2 < \dots < j_k$ refer to the chosen rows and columns, respectively, and set up the kth order matrix

$$\mathbf{A} = \begin{pmatrix} \alpha_{i_1 j_1} & \alpha_{i_1 j_2} & \dots & \alpha_{i_1 j_k} \\ \alpha_{i_2 j_1} & \alpha_{i_2 j_2} & \dots & \alpha_{i_2 j_k} \\ \dots & \dots & \dots & \dots \\ \alpha_{i_k j_1} & \alpha_{i_k j_2} & \dots & \alpha_{i_k j_k} \end{pmatrix}.$$

The determinant M_k of this matrix is called the *kth order minor of the matrix* **A**.

It is easy to see that an $m \times n$ matrix has minors of orders 1, 2, ..., min (m, n).

Example. We consider an 11 × 14 matrix shown in the upper part of Fig. 5.6. Let us choose 7 rows and 7 columns, say, the rows numbered 1, 3, 4, 6, 8, 9, 10 and the columns numbered 2, 5, 6, 7, 10, 12, 13. If we set up a matrix of order 7 comprising the elements which are contained in both the rows and columns chosen, we get the matrix shown in the lower part of Fig. 5.6 provided that the mutual arrangement of these elements in the original matrix is preserved in the new one. The determinant of the matrix thus obtained is the 7th order minor of the original matrix.

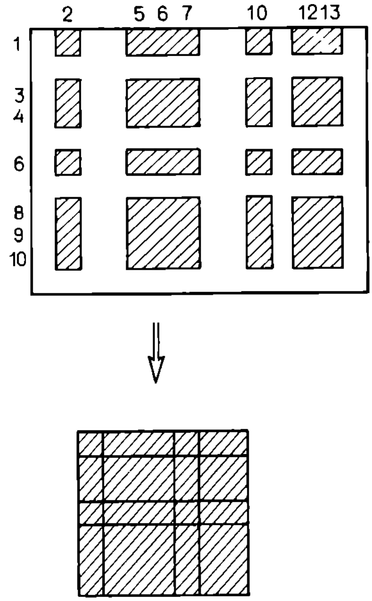


Fig. 5.6

Let **A** be a nonzero matrix. Then there exists a number r such that

(a) at least one rth order minor of **A** is distinct from zero and

(b) any sth order minor of **A** ($s > r$), if it exists, is equal to zero.

The number r is called the *rank of the matrix* **A**.

In symbols, we denote the rank of **A** by $r(\mathbf{A})$.

Obviously

$$0 \leqslant r(\mathbf{A}) \leqslant \min(m, n).$$

The minor M_r which is distinct from zero and such that its order is equal to the rank of the matrix **A** is called the *base minor* of **A**. The rows and columns of **A** comprising the elements of the base minor of **A** are called the *base rows* and the *base columns* of **A**.

Theorem 5.6. *Let* **A** *be an arbitrary matrix. Then*

(1) *The base rows (columns) of* **A** *are linearly independent.*

(2) *Each row (column) of* **A** *admits a representation in the form of a linear combination of the base rows (columns) of* **A**.

◄ For definiteness assume that the base minor of **A** is of order r and is in the top left corner of the matrix **A**, i.e.,

$$\mathbf{A} = \left(\begin{array}{ccc|cc} \alpha_{11} & \dots & \alpha_{1r} & \alpha_{1,r+1} & \alpha_{1n} \\ \dots & M_r & \dots & \dots & \dots \\ \alpha_{r1} & \dots & \alpha_{rr} & \alpha_{r,r+1} & \alpha_{rn} \\ \hline \alpha_{r+1,1} & \dots & \alpha_{r+1,r} & \alpha_{r+1,r+1} & \alpha_{r+1,n} \\ \dots & \dots & \dots & \dots & \dots \\ \alpha_{m1} & \dots & \alpha_{mr} & \alpha_{m,r+1} & \alpha_{mn} \end{array}\right). \quad (5.46)$$

Then the first r rows of **A** denoted by $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r$ become the base rows of **A**.

We shall prove that $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r$ are linearly independent. Let us suppose the opposite, i.e., the rows $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r$ are linearly dependent. Then one of the rows is a linear combination of the other $r - 1$ rows. Assume that

$$\mathbf{a}_r = \lambda_1\mathbf{a}_1 + \lambda_2\mathbf{a}_2 + \ldots + \lambda_{r-1}\mathbf{a}_{r-1}.$$

This means that the rth row of the base minor M_r is a linear combination of the other rows of M_r. Recall that any determinant such that one of its rows is a linear combination of the others is equal to zero. Hence, $M_r = 0$. But this contradicts the definition of the base minor. Consequently, the assumption on the linear dependence of $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r$ is false. Thus we infer that $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r$ are linearly independent.

Now we show that each row of **A** can be represented as a linear combination of the other base rows of **A**. First, we ascertain that for any i and j $(1 \leqslant i \leqslant m,\ 1 \leqslant j \leqslant n)$ there holds

$$\Delta = \left|\begin{array}{ccc|c} \alpha_{11} & \ldots & \alpha_{1r} & \alpha_{1j} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{r1} & \ldots & \alpha_{rr} & \alpha_{rj} \\ \hline \alpha_{i1} & \ldots & \alpha_{ir} & \alpha_{ij} \end{array}\right| = 0. \qquad (5.47)$$

Indeed, if $i \leqslant r$ Δ contains two identical rows, and if $j \leqslant r$ Δ contains two identical columns. Hence, $\Delta = 0$ in both these cases. When $i > r$ and $j > r$, Δ becomes the $(r + 1)$th order minor of the matrix **A**.

Let us fix a subscript i $(1 \leqslant i \leqslant m)$, the ith row, and expand the determinant Δ with respect to the last column. We get

$$\alpha_{1j}\Delta_1 + \alpha_{2j}\Delta_2 + \ldots + \alpha_{rj}\Delta_r + \alpha_{ij}M_r = 0. \qquad (5.48)$$

The identity (5.48) holds for any j $(1 \leqslant j \leqslant n)$ and $\Delta_1, \Delta_2, \ldots, \Delta_r$ are independent of j.

Setting

$$\lambda_1 = -\frac{\Delta_1}{M_r}, \quad \lambda_2 = -\frac{\Delta_2}{M_r}, \quad \ldots, \quad \lambda_r = -\frac{\Delta_r}{M_r},$$

we may write (5.48) in the form

$$\alpha_{ij} = \lambda_1\alpha_{1j} + \lambda_2\alpha_{2j} + \ldots + \lambda_r\alpha_{rj}, \quad j = 1, 2, \ldots, n,$$

or

$$\begin{array}{l} \alpha_{i1} = \lambda_1\alpha_{11} + \lambda_2\alpha_{21} + \ldots + \lambda_r\alpha_{r1} \\ \alpha_{i2} = \lambda_1\alpha_{12} + \lambda_2\alpha_{22} + \ldots + \lambda_r\alpha_{r2} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \alpha_{in} = \lambda_1\alpha_{1n} + \lambda_2\alpha_{2n} + \ldots + \lambda_r\alpha_{rn} \end{array} \qquad (5.49)$$

From (5.49) it follows that

$$\mathbf{a}_i = \lambda_1\mathbf{a}_1 + \lambda_2\mathbf{a}_2 + \ldots + \lambda_r\mathbf{a}_r.$$

Whence we conclude that the ith row of $\mathbf{A}$ is a linear combination of the base rows $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r$. Since the above reasoning is applicable to any other row of $\mathbf{A}$ we infer that each row of $\mathbf{A}$ is a linear combination of the base rows in $\mathbf{A}$. ►

Proposition. If a matrix is subjected to elementary operations its rank does not increase.

◄ Let $\tilde{\mathbf{A}}$ be a matrix of rank $\tilde{r}$ obtained from a matrix $\mathbf{A}$ of rank r by interchanging its rows. We consider an arbitrary minor $\tilde{M}_s$ of order s in $\tilde{\mathbf{A}}$ and choose in $\mathbf{A}$ a minor M_s of the same order s, such that the elements of M_s are arranged in the rows and columns of $\mathbf{A}$ in the same manner as the rows and columns of $\tilde{\mathbf{A}}$ comprising the elements of $\tilde{M}_s$. Observe that interchanging the rows of $\mathbf{A}$ rearranges the rows without altering them. This means that the rows in M_s may only differ from those in $\tilde{M}_s$ by their arrangements. Whence it follows that either $\tilde{M}_s = M_s$ or $\tilde{M}_s = -M_s$.

By definition of the rank all minors of $\mathbf{A}$ whose orders exceed r are equal to zero. From the above consideration we infer that any sth ($s > r$) order minor $\tilde{M}_s$ of $\tilde{\mathbf{A}}$ is equal to zero, i.e., $\tilde{M}_s = 0$. This means that the rank of $\tilde{\mathbf{A}}$ cannot exceed the rank of $\mathbf{A}$, i.e., $\tilde{r} \leqslant r$.

Similarly, we may show that $\tilde{r} \leqslant r$ in the cases when the matrix $\tilde{\mathbf{A}}$ is obtained from $\mathbf{A}$ by elementary operations of the other two types. Obviously, this assertion is true for elementary column operations. ►

Theorem 5.7. *If an arbitrary matrix* $\mathbf{A}$ *is subjected to elementary row (column) operations its rank remains unchanged.*

◄ It suffices to recall that if the matrix $\tilde{\mathbf{A}}$ is obtained from the matrix $\mathbf{A}$ by elementary operations, the matrix $\mathbf{A}$ is obtainable from $\tilde{\mathbf{A}}$ by elementary operations of the same type. Taking into account the above assertion, we conclude that $r \leqslant \tilde{r}$.

Comparing the inequalities $\tilde{r} \leqslant r$ and $r \leqslant \tilde{r}$, we arrive at the desired identity

$$\tilde{r} = r. \quad ►$$

Remark. The number of nonzero rows in a matrix of schematic form is equal to the rank of the matrix.

Indeed, for a matrix of schematic form the rth order minor whose elements are contained in the first r rows and in the columns $k_1, k_2, \ldots, k_r$ is distinct from zero, i.e.,

$$\begin{vmatrix} \alpha_{1k_1}^{(1)} & & 0 \\ & \alpha_{2k_2}^{(2)} & \\ & & \ddots & \\ 0 & & & \alpha_{rk_r}^{(r)} \end{vmatrix} = \alpha_{1k_1}^{(1)}\alpha_{2k_2}^{(2)} \dots \alpha_{rk_r}^{(r)} \neq 0$$

and any minor of order s ($s > r$) comprises the zero row and, consequently, is equal to zero.

Thus, subjecting a matrix to a sequence of elementary operations, we may easily and effectively define the rank of this matrix.

Example. Find the rank of

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 3 & -2 & 4 \\ 4 & -2 & 5 & 1 & 7 \\ 2 & -1 & 1 & 8 & 2 \end{pmatrix}.$$

◄ *Step 1.* Subtracting the multiples of the first row by 2 and 1 from the second and third rows, respectively, we get

$$r(\mathbf{A}) = r\left(\begin{array}{cc|ccc} 2 & -1 & 3 & -2 & 4 \\ \cline{1-2} 0 & 0 & -1 & 5 & -1 \\ 0 & 0 & -2 & 10 & -2 \end{array}\right).$$

Step 2. Subtracting the second row multiplied by 2 from the third row, we get

$$r(\mathbf{A}) = r\left(\begin{array}{cc|ccc} 2 & -1 & 3 & -2 & 4 \\ \cline{1-2} 0 & 0 & -1 & 5 & -1 \\ \cline{3-5} 0 & 0 & 0 & 0 & 0 \end{array}\right) = 2. \text{ ►}$$

5.5 Systems of Linear Equations

Definitions. Suppose that we are given a matrix

$$\hat{\mathbf{A}} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} & \beta_1 \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} & \beta_2 \\ \dots & \dots & \dots & \dots & \dots \\ \alpha_{m1} & \alpha_{m2} & \dots & \alpha_{mn} & \beta_m \end{pmatrix}, \tag{5.50}$$

whose first n columns are nonzero.

A collection of relations of the form

$$\begin{gathered} \alpha_{11}x_1 + \alpha_{12}x_2 + \dots + \alpha_{1n}x_n = \beta_1, \\ \alpha_{21}x_2 + \alpha_{22}x_2 + \dots + \alpha_{2n}x_n = \beta_2, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \alpha_{m1}x_1 + \alpha_{m2}x_2 + \dots + \alpha_{mn}x_n = \beta_m, \end{gathered} \tag{5.51}$$

where $x_1, x_2, \ldots, x_n$ are quantities to be defined (*unknowns*), is called the *system of m linear equations in n unknowns* or the *linear system.*

The numbers α_{ij} $(i = 1, 2, \ldots, m; j = 1, 2, \ldots, n)$ are called the *coefficients of the linear system* (5.51) and the numbers β_i $(i = 1, 2, \ldots, m)$ are called the *constant* or *free terms.*

A *solution* of the linear system (5.51) is an ordered n-tuple of numbers $\gamma_1, \gamma_2, \ldots, \gamma_n$ which, being substituted into (5.51) for the unknowns $x_1, x_2, \ldots, x_n$, turns each equation of (5.51) into identity.

The system of linear equations is said to be *consistent* if it has a solution and *inconsistent* if it has no solutions.

Two solutions $\gamma_1, \gamma_2, \ldots, \gamma_n$ and $\gamma_1', \gamma_2', \ldots, \gamma_n'$ are called *distinct* if at least one of the identities $\gamma_1 = \gamma_1', \gamma_2 = \gamma_2', \ldots, \gamma_n = \gamma_n'$ is violated.

A consistent system is called *determinate* if it has a unique solution, and *indeterminate* if it has at least two distinct solutions.

In matrix form, the system (5.91) may be written as the *matrix* equation

$$\mathbf{AX} = \mathbf{b}, \tag{5.52}$$

where

$$\mathbf{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{m1} & \alpha_{m2} & \ldots & \alpha_{mn} \end{pmatrix},$$

$$\mathbf{b} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}. \tag{5.53}$$

The matrix $\mathbf{A}$ is called the *coefficient matrix* of the system (5.51), the column-vector $\mathbf{b}$ is called the *column-vector of constant terms* and the column-vector $\mathbf{X}$ is called the *column-vector of unknowns.* The matrix

$$\hat{\mathbf{A}} = (\mathbf{A} \,|\, \mathbf{b})$$

is called the *augmented matrix* of the system (5.51). The solution of the matrix equation (5.52) is the column-vector

$$\mathbf{\Gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_n \end{pmatrix}.$$

Theorem 5.8 (Kronecker-Capelli theorem). *The system of linear equations (the linear system) is consistent if and only if the rank of the coefficient matrix is equal to the rank of the augmented matrix.*

◄ Suppose that the system (5.51) is consistent. This means that the n-tuple of numbers $\gamma_1, \gamma_2, \ldots, \gamma_n$ turns each equation of (5.51) into identity, that is,

$$\begin{aligned} \alpha_{11}\gamma_1 &+ \alpha_{12}\gamma_2 + \ldots + \alpha_{1n}\gamma_n = \beta_1 \\ \alpha_{21}\gamma_1 &+ \alpha_{22}\gamma_2 + \ldots + \alpha_{2n}\gamma_n = \beta_2 \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \alpha_{m1}\gamma_1 &+ \alpha_{m2}\gamma_2 + \ldots + \alpha_{mn}\gamma_n = \beta_m \end{aligned}$$

From the above identities it follows that in the augmented matrix $\hat{\mathbf{A}} = (\mathbf{A} \mid \mathbf{b})$ the column-vector $\mathbf{b}$ of constant terms is a linear combination of the other columns in $\hat{\mathbf{A}}$, i.e., $\mathbf{b}$ is a linear combination of the columns of $\mathbf{A}$.

Adding to the column-vector $\mathbf{b}$ the first column of $\mathbf{A}$ multiplied by $-\gamma_1$, the second column of $\mathbf{A}$ multiplied by $-\gamma_2$, etc., and finally the nth column of $\mathbf{A}$ multiplied by $-\gamma_n$, we reduce the augmented matrix $\hat{\mathbf{A}}$ to the matrix

$$\tilde{\mathbf{A}} = (\mathbf{A} \mid \mathbf{0}).$$

The rank of $\tilde{\mathbf{A}}$ is equal to that of $\hat{\mathbf{A}}$ since, by Theorem 5.7, the elementary column operations leave the rank unchanged. On the other hand, it is clear that the rank of $\tilde{\mathbf{A}} = (\mathbf{A} \mid \mathbf{0})$ is equal to that of $\mathbf{A}$. Whence we infer that $r(\hat{\mathbf{A}}) = r(\tilde{\mathbf{A}}) = r(\mathbf{A})$.

Now we suppose that the rank of the coefficient matrix is equal to that of the augmented matrix.

Since $\hat{\mathbf{A}} = (\mathbf{A} \mid \mathbf{b})$, the matrices $\mathbf{A}$ and $\hat{\mathbf{A}}$ have common base minors. Let the common base minor be of order r and let this base minor be in the top left corners of the respective matrices.

By Theorem 5.6 any column in $\hat{\mathbf{A}}$ admits a representation in the form of a linear combination of the base columns of $\hat{\mathbf{A}}$. In particular, the column-vector of constant terms, i.e., the $(n + 1)$th column in $\hat{\mathbf{A}}$, may be written as

$$\gamma_1 \begin{pmatrix} \alpha_{11} \\ \alpha_{21} \\ \vdots \\ \alpha_{m1} \end{pmatrix} + \gamma_2 \begin{pmatrix} \alpha_{12} \\ \alpha_{22} \\ \vdots \\ \alpha_{m2} \end{pmatrix} + \ldots + \gamma_r \begin{pmatrix} \alpha_{1r} \\ \alpha_{2r} \\ \vdots \\ \alpha_{mr} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix}$$

or

$$\begin{aligned} \alpha_{11}\gamma_1 &+ \alpha_{12}\gamma_2 + \ldots + \alpha_{1r}\gamma_r = \beta_1, \\ \alpha_{21}\gamma_1 &+ \alpha_{22}\gamma_2 + \ldots + \alpha_{2r}\gamma_r = \beta_2, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \alpha_{m1}\gamma_1 &+ \alpha_{m2}\gamma_2 + \ldots + \alpha_{mr}\gamma_r = \beta_m. \end{aligned}$$

It is easy to see that the ordered n-tuple of numbers

$$\gamma_1, \gamma_2, \ldots, \gamma_r, \quad 0, \ldots, 0$$

turns each equation of the original system into identity. Whence it follows that the system (5.51) is consistent. ►

Equivalence of linear systems. We shall call a collection of all solutions of a given linear system a *set of solutions*.

Two linear systems comprising equal numbers of unknowns are said to be *equivalent* if their sets of solutions coincide. In other words, two linear systems are equivalent if each solution of one system is a solution of the other one and vice versa, or if both these systems have no solutions.

It is clear that any linear system is uniquely determined by its augmented matrix. Suppose that $\hat{\mathbf{A}}$ and $\hat{\mathbf{A}}'$ are two $m \times (n+1)$ augmented matrices of the linear systems

$$\begin{aligned} \alpha_{11}x_1 + \alpha_{12}x_2 + \ldots + \alpha_{1n}x_n &= \beta_1, \\ \alpha_{21}x_1 + \alpha_{22}x_2 + \ldots + \alpha_{2n}x_n &= \beta_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \alpha_{m1}x_1 + \alpha_{m2}x_2 + \ldots + \alpha_{mn}x_n &= \beta_m \end{aligned} \tag{$*$}$$

and

$$\begin{aligned} \alpha'_{11}x_1 + \alpha'_{12}x_2 + \ldots + \alpha'_{1n}x_n &= \beta'_1, \\ \alpha'_{21}x_1 + \alpha'_{22}x_2 + \ldots + \alpha'_{2n}x_n &= \beta'_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \alpha'_{m1}x_1 + \alpha'_{m2}x_2 + \ldots + \alpha'_{mn}x_n &= \beta'_m. \end{aligned} \tag{$*'$}$$

The system (*′) is said to be obtained from the system (*) by elementary operations if the augmented matrix $\hat{\mathbf{A}}'$ is obtained from the augmented matrix $\hat{\mathbf{A}}$ by elementary row operations.

Theorem 5.9. *If the linear system (*′) is obtained from the linear system (*) by elementary operations these systems are equivalent.*

◄ First, we suppose that the system (*) is consistent. Let $\gamma_1, \gamma_2, \ldots, \gamma_n$ be a solution of (*). We show that this n-tuple of numbers turns each equation of (*′) into identity. Consider three cases which correspond to elementary operations of different kinds:

(a) Assume that the system (*′) is obtained from (*) by interchanging rows in $\hat{\mathbf{A}}$. This means that the matrix $\hat{\mathbf{A}}'$ comprises the same collection of rows as $\hat{\mathbf{A}}$ and, consequently, the system (*′) contains the same collection of equations as (*). Whence we conclude that if the n-tuple $\gamma_1, \gamma_2, \ldots, \gamma_n$ turns each equation of (*) into identity, it also turns each equation of (*′) into identity.

(b) Assume that the system (*′) is obtained from (*) by multiplying the kth row in $\hat{\mathbf{A}}$ by a nonzero scalar λ. Then the kth equation of (*′)

will be of the form

$$(\lambda\alpha_{k1})x_1 + (\lambda\alpha_{k2})x_2 + \ldots + (\lambda\alpha_{kn})x_n = \lambda\beta_k$$

and the other $(m - 1)$ equations will be the same as in $(*)$.

Substituting $\gamma_1, \gamma_2, \ldots, \gamma_n$ into the kth equation of $(*)$, we obtain the identity

$$\alpha_{k1}\gamma_1 + \alpha_{k2}\gamma_2 + \ldots + \alpha_{kn}\gamma_n = \beta_k.$$

Multiplying this identity by $\lambda \neq 0$, we arrive at

$$(\lambda\alpha_{k1})\gamma_1 + (\lambda\alpha_{k2})\gamma_2 + \ldots + (\lambda\alpha_{kn})\gamma_n = \lambda\beta_k.$$

Whence we conclude that the solution of $(*)$ is also the solution of $(*')$ since $\gamma_1, \gamma_2, \ldots, \gamma_n$ turns the kth and all other equations of $(*')$ into identities.

(c) Now we assume that the system $(*')$ is obtained by adding to the lth row in $\hat{\mathbf{A}}$ the kth row multiplied by a scalar μ. Then the lth equation of $(*')$ will be of the form

$$(\alpha_{l1} + \mu\alpha_{k1})x_1 + (\alpha_{l2} + \mu\alpha_{k2})x_2 + \ldots + (\alpha_{ln} + \mu\alpha_{kn})x_n = \beta_l + \mu\beta_k \tag{5.54}$$

and the other $(m - 1)$ equations in $(*')$ will be the same as in $(*)$.

Substituting $\gamma_1, \gamma_2, \ldots, \gamma_n$ into the kth and lth equations of $(*)$, we obtain two identities

$$\alpha_{k1}\gamma_1 + \alpha_{k2}\gamma_2 + \ldots + \alpha_{kn}\gamma_n = \beta_k,$$
$$\alpha_{l1}\gamma_1 + \alpha_{l2}\gamma_2 + \ldots + \alpha_{ln}\gamma_n = \beta_l.$$

Substituting the multiple of the first identity by μ into the second one, we obtain

$$(\alpha_{l1} + \mu\alpha_{k1})\gamma_1 + (\alpha_{l2} + \mu\alpha_{k2})\gamma_2 + \ldots + (\alpha_{ln} + \mu\alpha_{kn})\gamma_n = \beta_l + \mu\beta_k.$$

Whence we infer that the solution $\gamma_1, \gamma_2, \ldots, \gamma_n$ of $(*)$ is also the solution of $(*')$. Notice that by substituting $\gamma_1, \gamma_2, \ldots, \gamma_n$ into (5.54), we reach the same result.

Thus, in each case any solution of the system $(*)$ is a solution of the system $(*')$.

Observe that the system $(*)$ may also be obtained from the system $(*')$ by elementary operations. Repeating the previous arguments, we infer that any solution of the system $(*')$ is a solution of the system $(*)$.

Second, we suppose that the system $(*)$ is inconsistent. Then the system $(*')$ is also inconsistent. To prove this fact, assume the opposite, i.e., assume that the system $(*')$ is consistent. Then Theorem 5.9 implies that the system

(∗) is also consistent. However, this contradicts to what we have assumed. Whence we conclude that if the system (∗) is inconsistent, so is the system (∗′). ►

Remark. Clearly, if the system (∗′) is obtained from the system (∗) by a finite sequence of elementary operations, then these systems are equivalent.

Gaussian elimination. When we wish to solve a linear system we should, first, find out whether the system is consistent, and if this is the case, second, find the set of its solutions.

We shall outline a method of solving a given linear system called the *method of Gaussian elimination.* The method employs elementary operations to reduce a given linear system to a system of simpler form whose solutions are easily computed.

The elementary operations on a linear system are equivalent to the elementary row operations on its augmented matrix. So we shall simultaneously consider the linear system

$$\begin{array}{l} \alpha_{11}x_1 + \alpha_{12}x_2 + \ldots + \alpha_{1n}x_n = \beta_1, \\ \alpha_{21}x_1 + \alpha_{22}x_2 + \ldots + \alpha_{2n}x_n = \beta_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \alpha_{m1}x_1 + \alpha_{m2}x_2 + \ldots + \alpha_{mn}x_n = \beta_m \end{array} \qquad (*)$$

and the augmented matrix

$$\hat{\mathbf{A}} = \left(\begin{array}{cccc|c} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} & \beta_1 \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} & \beta_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \alpha_{m1} & \alpha_{m2} & \ldots & \alpha_{mn} & \beta_m \end{array}\right).$$

In the previous sections we have shown that the elementary row operations may reduce the augmented matrix $\hat{\mathbf{A}}$ to schematic form

$$\hat{\mathbf{A}}' = \left(\begin{array}{cccc|c} \alpha^{(1)}_{1k_1} & & & & \\ & \alpha^{(2)}_{2k_2} & & & \\ & & \ddots & & \\ & & & \alpha^{(r)}_{rk_r} & \\ & & & & \beta^{(r)}_{r+1} \\ & & & & 0 \\ & 0 & & & \cdots \\ & & & & 0 \end{array}\right).$$

Similarly, the system (∗) is reduced to the form with the augmented matrix $\hat{\mathbf{A}}$.

If the constant term $\beta_{r+1}^{(r)}$ is distinct from zero the new linear system (and, consequently, the original one) is inconsistent. Indeed, the $(r + 1)$th equation takes the form

$$0 \cdot x_1 + 0 \cdot x_2 + \ldots + 0 \cdot x_n = \beta_{r+1}^{(r)} \neq 0.$$

This means that there exists no n-tuple of numbers turning the above equation into identity.

Let $\beta_{r+1}^{(r)} = 0$. Then only the first r rows in $\hat{\mathbf{A}}'$ are distinct from zero. To simplify the notation we set

$$k_2 = 2,\ k_3 = 3,\ \ldots,\ k_r = r.$$

(This may be made by renumbering $y_1 = x_1,\ y_2 = x_{k_2},\ \ldots,\ y_r = x_{k_r},\ \ldots$.) The linear system takes the form

$$\begin{aligned} \alpha_{11}^{(1)}x_1 + \alpha_{12}^{(1)}x_2 + \ldots + \alpha_{1r}^{(1)}x_r + \ldots + \alpha_{1n}^{(1)}x_n &= \beta_1^{(1)}, \\ \alpha_{22}^{(2)}x_2 + \ldots + \alpha_{2r}^{(2)}x_r + \ldots + \alpha_{2n}^{(2)}x_n &= \beta_2^{(2)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ \alpha_{rr}^{(r)}x_r + \ldots + \alpha_{rn}^{(r)}x_n &= \beta_r^{(r)}, \end{aligned} \qquad (*')$$

where $\alpha_{11}^{(1)} \neq 0,\ \alpha_{22}^{(2)} \neq 0,\ \ldots,\ \alpha_{rr}^{(r)} \neq 0$. Then

(1) The number n of unknowns in $(*')$ is equal to the number r of equations, i.e., $r = n$. Then the system $(*')$ takes the form

$$\begin{aligned} \alpha_{11}^{(1)}x_1 + \alpha_{12}^{(1)}x_2 + \ldots + \alpha_{1,n-1}^{(1)}x_{n-1} + \alpha_{1n}^{(1)}x_n &= \beta_1^{(1)}, \\ \alpha_{22}^{(2)}x_2 + \ldots + \alpha_{2,n-1}^{(2)}x_{n-1} + \alpha_{2n}^{(2)}x_n &= \beta_2^{(2)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ \alpha_{n-1,n-1}^{(n-1)}x_{n-1} + \alpha_{n-1,n}^{(n-1)}x_n &= \beta_{n-1}^{(n-1)}, \\ \alpha_{nn}^{(n)}x_n &= \beta_n^{(n)}, \end{aligned}$$

where $\alpha_{kk}^{(k)} \neq 0,\ k = 1, 2, \ldots$.

The last equation defines uniquely the value of x_n. Substituting the value of x_n into the $(n - 1)$th equation, we get the value of x_{n-1}. Then substituting the values of x_n and x_{n-1} into the $(n - 2)$th equation, we obtain the value of x_{n-2}. Continuing this process, we obtain the values of $x_{n-3},\ x_{n-4},\ \ldots,\ x_3,\ x_2$. Finally, substituting the values of $x_2,\ x_3,\ \ldots,\ x_n$ into the first equation we obtain the value of x_1.

Thus the system $(*')$ has a unique solution and so does the system $(*)$.

(2) The number n of unknowns in $(*')$ is greater than the number r of equations, i.e., $n > r$.

We set the unknowns $x_{r+1},\ x_{r+2},\ \ldots,\ x_n$ equal to arbitrary numbers $\gamma_{r+1},\ \gamma_{r+2},\ \ldots,\ \gamma_n$. Then transposing the corresponding terms to the right, we have the system

$$\begin{aligned}
\alpha_{11}^{(1)}x_1 + \alpha_{12}^{(1)}x_2 + \dots + \alpha_{1r}^{(1)}x_r &= \beta_1^{(1)} - \alpha_{1,r+1}^{(1)}\gamma_{r+1} - \dots - \alpha_{1n}^{(1)}\gamma_n, \\
\alpha_{22}^{(2)}x_2 + \dots + \alpha_{2r}^{(2)}x_r &= \beta_2^{(2)} - \alpha_{2,r+1}^{(2)}\gamma_{r+1} - \dots - \alpha_{2n}^{(2)}\gamma_n, \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\alpha_{rr}^{(r)}x_r &= \beta_r^{(r)} - \alpha_{r,r+1}^{(r)}\gamma_{r+1} - \dots - \alpha_{rn}^{(r)}\gamma_n.
\end{aligned}$$

Similarly to the case (1) we obtain the values of $x_r, x_{r-1}, \dots, x_2, x_1$. Since $\gamma_{r+1}, \gamma_{r+2}, \dots, \gamma_n$ are arbitrary numbers we infer that the system has infinitely many solutions. ►

Example. Solve the system

$$\begin{cases} 3x_1 - 5x_2 + 2x_3 + 4x_4 = 2, \\ 7x_1 - 4x_2 + x_3 + 3x_4 = 5, \\ 5x_1 + 7x_2 - 4x_3 - 6x_4 = 3. \end{cases}$$

◄ Let us arrange the augmented matrix

$$\hat{\mathbf{A}} = \left(\begin{array}{cccc|c} 3 & -5 & 2 & 4 & 2 \\ 7 & -4 & 1 & 3 & 5 \\ 5 & 7 & -4 & -6 & 3 \end{array}\right)$$

and reduce **A** to schematic form.

Step 1. To get the matrix with $\alpha_{11} = 1$ we subtract the first row multiplied by 2 from the second row and interchange the first row and the obtained second one. We have

$$\hat{\mathbf{A}} = \left(\begin{array}{cccc|c} 1 & 6 & -3 & -5 & 1 \\ 3 & -5 & 2 & 4 & 2 \\ 5 & 7 & -4 & -6 & 3 \end{array}\right).$$

Subtracting the multiples of the first row by 3 and 5 from the second and third rows, respectively, we get

$$\left(\begin{array}{c|cccc|c} \multicolumn{1}{c}{1} & 6 & -3 & -5 & & 1 \\ \cline{1-1} 0 & -23 & 11 & 19 & & -1 \\ 0 & -23 & 11 & 19 & & -2 \end{array}\right).$$

Step 2. Subtracting the second row from the third one, we have

$$\left(\begin{array}{c|cccc|c} \multicolumn{1}{c}{1} & 6 & -3 & -5 & & 1 \\ \cline{1-1} 0 & -23 & 11 & 19 & & -1 \\ \cline{2-4} \multicolumn{1}{c}{0} & 0 & 0 & 0 & & -1 \end{array}\right).$$

Since the rank of the coefficient matrix is equal to 2 and the rank of the augmented matrix is equal to 3 we conclude that the linear system is inconsistent. ►

Example. Solve the system

$$\begin{cases} 2x_1 + 5x_2 - 8x_3 = 8, \\ 4x_1 + 3x_2 - 9x_3 = 9, \\ 2x_1 + 3x_2 - 5x_3 = 7, \\ x_1 + 8x_2 - 7x_3 = 12. \end{cases}$$

◀ Arrange the augmented matrix

$$\hat{\mathbf{A}} = \left(\begin{array}{ccc|c} 2 & 5 & -8 & 8 \\ 4 & 3 & -9 & 9 \\ 2 & 3 & -5 & 7 \\ 1 & 8 & -7 & 12 \end{array}\right).$$

Step 1. Interchanging the first and fourth rows, we have

$$\left(\begin{array}{ccc|c} 1 & 8 & -7 & 12 \\ 4 & 3 & -9 & 9 \\ 2 & 3 & -5 & 7 \\ 2 & 5 & -8 & 8 \end{array}\right).$$

Subtracting the multiples of the first row by 4, 2 and 2 from the second, third and fourth rows, respectively, we get

$$\left(\begin{array}{c|cc|c} \multicolumn{1}{c}{1} & 8 & -7 & 12 \\ \cline{1-1} 0 & -29 & 19 & -39 \\ 0 & -13 & 9 & -17 \\ 0 & -11 & 6 & -16 \end{array}\right).$$

Step 2. To simplify computations we subtract the third row multiplied by 2 from the second row and then the fourth row from the third one. We have

$$\left(\begin{array}{c|cc|c} \multicolumn{1}{c}{1} & 8 & -7 & 12 \\ \cline{1-1} 0 & -3 & 1 & -5 \\ 0 & -2 & 3 & -1 \\ 0 & -11 & 6 & -16 \end{array}\right).$$

Subtracting the third row from the second one, we get

$$\left(\begin{array}{c|cc|c} \multicolumn{1}{c}{1} & 8 & -7 & 12 \\ \cline{1-1} 0 & -1 & -2 & -4 \\ 0 & -2 & 3 & -1 \\ 0 & -11 & 6 & -16 \end{array}\right).$$

Subtracting the multiples of the second rows by 2 and 11 from the third and fourth rows, respectively, we obtain

$$\left(\begin{array}{ccc|c} 1 & 8 & -7 & 12 \\ 0 & -1 & -2 & -4 \\ 0 & 0 & 7 & 7 \\ 0 & 0 & 28 & 28 \end{array}\right).$$

Step 3. Subtracting the third row multiplied by 4 from the fourth row, we have

$$\left(\begin{array}{ccc|c} 1 & 8 & -7 & 12 \\ 0 & -1 & -2 & -4 \\ 0 & 0 & 7 & 7 \\ 0 & 0 & 0 & 0 \end{array}\right).$$

Multiplying the third row by 1/7, we obtain

$$\left(\begin{array}{ccc|c} 1 & 8 & -7 & 12 \\ 0 & -1 & -2 & -4 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right).$$

The system is consistent since the rank of the coefficient matrix is equal to the rank of the augmented matrix, that is, $r(\mathbf{A}) = r(\hat{\mathbf{A}}) = 3$. Since the rank is equal to the number of unknowns the system has a unique solution.

Thus the original system is equivalent to the system

$$\begin{cases} x_1 + 8x_2 - 7x_3 = 12, \\ \quad\; - x_2 - 2x_3 = -4, \\ \qquad\qquad\;\; x_3 = 1. \end{cases}$$

From the third equation we have $x_3 = 1$. Substituting $x_3 = 1$ into the second equation, we have $-x_2 - 2 = -4$. Whence, $x_2 = 2$. Substituting the values of x_2 and x_3 into the first equation, we get $x_1 + 16 - 7 = 12$. Whence $x_1 = 3$. Thus the system has the unique solution $x_1 = 3$, $x_2 = 2$, $x_3 = 1$. ►

Example. Solve the system

$$\begin{cases} 3x_1 - 2x_2 + 5x_3 + 4x_4 = 2, \\ 6x_1 - 4x_2 + 4x_3 + 3x_4 = 3, \\ 9x_1 - 6x_2 + 3x_3 + 2x_4 = 4. \end{cases}$$

◄ Arrange the augmented matrix

$$\left(\begin{array}{cccc|c} 3 & -2 & 5 & 4 & 2 \\ 6 & -4 & 4 & 3 & 3 \\ 9 & -6 & 3 & 2 & 4 \end{array}\right).$$

Step 1. Subtracting the multiples of the first row by 2 and 3 from the second and third rows, respectively, we have

$$\left(\begin{array}{cccc|c} 3 & -2 & 5 & 4 & 2 \\ \hline 0 & 0 & -6 & -5 & -1 \\ 0 & 0 & -12 & -10 & -2 \end{array}\right).$$

Step 2. Subtracting the multiple of the second row by 2 from the third row, we get

$$\left(\begin{array}{cccc|c} 3 & -2 & 5 & 4 & 2 \\ \hline 0 & 0 & -6 & -5 & -1 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

The system is consistent ($r(\mathbf{A}) = r(\hat{\mathbf{A}}) = 2$) and has infinitely many solutions since the rank of the coefficient matrix is smaller than the number of unknowns ($r(\mathbf{A}) < 4$).

The original system is equivalent to the system

$$\begin{cases} 3x_1 - 2x_2 + 5x_3 + 4x_4 = 2, \\ \qquad\qquad\; - 6x_3 - 5x_4 = -1. \end{cases}$$

Let us find the solution of this system.

Set x_2 and x_4 equal to arbitrary numbers γ_2 and γ_4. Then transposing the corresponding terms of the system to the right, we have

$$\begin{cases} 3x_1 + 5x_3 = 2 + 2\gamma_2 - 4\gamma_4, \\ \qquad\; 6x_3 = 1 \qquad\quad - 5\gamma_4. \end{cases}$$

From the last equation we get

$$x_3 = \frac{1}{6}(1 - 5\gamma_4),$$

where γ_4 is an arbitrary number.

Substituting the expression for x_3 into the first equation, we obtain

$$x_1 = \frac{1}{18}(7 + 12\gamma_2 + \gamma_4),$$

where γ_2 and γ_4 are arbitrary numbers.

The general solution of the system is

$$x_1 = \frac{1}{18}(7 + 12\gamma_2 + \gamma_4), \quad x_2 = \gamma_2, \quad x_3 = \frac{1}{6}(1 - 5\gamma_4), \quad x_4 = \gamma_4.$$

Any specific solution is obtainable from the general one by giving specific values to γ_2 and γ_4. For instance, if we set $x_2 = \gamma_2 = 1$ and $x_4 = \gamma_4 = -1$, we get $x_1 = x_3 = 1$. Thus the specific solution of the system is

$$x_1 = 1, \; x_2 = 1, \; x_3 = 1, \; x_4 = -1. \blacktriangleright$$

Cramer's rule. Suppose we are given a system of n linear equations in n unknowns

$$\begin{cases} \alpha_{11}x_1 + \alpha_{12}x_2 + \ldots + \alpha_{1n}x_n = \beta_1, \\ \alpha_{21}x_1 + \alpha_{22}x_2 + \ldots + \alpha_{2n}x_n = \beta_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \alpha_{n1}x_1 + \alpha_{n2}x_2 + \ldots + \alpha_{nn}x_n = \beta_n \end{cases} \tag{5.55}$$

or, in matrix form,

$$\mathbf{AX} = \mathbf{b}. \tag{5.56}$$

System (5.55) is called the *quadratic system* of linear equations.

If the square matrix $\mathbf{A}$ is nonsingular the system (5.55) is consistent and has a unique solution since $r(\mathbf{A}) = n$.

Pre-multiplying (5.56) by the matrix $\mathbf{A}^{-1}$ inverse to $\mathbf{A}$, we have

$$\mathbf{X} = \mathbf{A}^{-1}\mathbf{b}$$

or, using formula (5.39),

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \frac{1}{|\mathbf{A}|}\begin{pmatrix} A_{11} & A_{21} & \ldots & A_{n1} \\ A_{12} & A_{22} & \ldots & A_{n2} \\ \ldots & \ldots & \ldots & \ldots \\ A_{1n} & A_{2n} & & A_{nn} \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}.$$

After a little algebra in the right-hand side of the above expression, we have

$$x_j = \frac{1}{|\mathbf{A}|}\sum_{k=1}^{n} \beta_k A_{kj}, \quad j = 1, 2, \ldots, n$$

or, as quotient of determinants,

$$x_j = \frac{\begin{vmatrix} \alpha_{11} & \alpha_{12} \ldots & \overset{j}{\beta_1} & & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} \ldots & \beta_2 & \ldots & \alpha_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \alpha_{n1} & \alpha_{n2} \ldots & \beta_n & \ldots & \alpha_{nn} \end{vmatrix}}{\begin{vmatrix} \alpha_{11} & \alpha_{12} \ldots & \alpha_{1j} & & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} \ldots & \alpha_{2j} & \ldots & \alpha_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \alpha_{n1} & \alpha_{n2} \ldots & \alpha_{nj} & \ldots & \alpha_{nn} \end{vmatrix}}, \quad j = 1, 2, \ldots, n, \tag{5.57}$$

where the numerator is the determinant of the matrix obtained from $\mathbf{A}$ by substituting the column-vector $\mathbf{b}$ for the jth column and the denominator is the determinant of the coefficient matrix $\mathbf{A}$.

It is important to mention that Cramer's rule (5.57) is mainly of theoretical interest and is of no use for practical purposes it being awkward to solve linear systems except those in two or three unknowns.

Remark. Cramer's rule involves $n + 1$ determinants of order n thus requiring a vast amount of direct computations which is proportional to $n!n$ and far exceeds the number of arithmetic operations required by the Gaussian elimination. When $n = 30$ the task of computing the solution of the linear system becomes time consuming even for advanced computers. The total number of arithmetic operations involved in Gaussian elimination is proportional to n^3.

To solve the quadratic linear system

$$\mathbf{AX} = \mathbf{b},$$

where $\mathbf{A}$ is a nonsingular square matrix, we should proceed as follows:

Step 1. Arrange the augmented matrix

$$\hat{\mathbf{A}} = (\mathbf{A} \mid \mathbf{b}) = \left(\begin{array}{cccc|c} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} & \beta_1 \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} & \beta_2 \\ \dots & \dots & \dots & \dots & \dots \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} & \beta_n \end{array}\right).$$

Step 2. By elementary row operations reduce $\hat{\mathbf{A}}$ to the form

$$(\tilde{\mathbf{A}} \mid \tilde{\mathbf{b}}) = \left(\begin{array}{cccc|c} 1 & \tilde{\alpha}_{12} & \dots & \tilde{\alpha}_{1n} & \tilde{\beta}_1 \\ 0 & 1 & \dots & \tilde{\alpha}_{2n} & \tilde{\beta}_2 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & \tilde{\beta}_n \end{array}\right),$$

where $\tilde{\mathbf{A}}$ is a triangular matrix.

Step 3. By elementary row operations reduce $(\tilde{\mathbf{A}} \mid \tilde{\mathbf{b}})$ to the form

$$(\mathbf{I} \mid \mathbf{c}) = \left(\begin{array}{cccc|c} 1 & 0 & \dots & 0 & \gamma_1 \\ 0 & 1 & \dots & 0 & \gamma_2 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & \dots & 1 & \gamma_n \end{array}\right),$$

where I is an identity matrix.

Step 4. Write down the system corresponding to the augmented matrix $(\mathbf{I} \mid \mathbf{c})$

$$\begin{array}{llll} x_1 & & & = \gamma_1 \\ & x_2 & & = \gamma_2 \\ \dots & \dots & \dots & \dots \\ & & x_n & = \gamma_n \end{array}$$

The n-tuple

$$x_1 = \gamma_1, \quad x_2 = \gamma_2, \quad \dots, \quad x_n = \gamma_n$$

is the solution of the original system.

Homogeneous linear systems. The system of m linear equations in n unknowns is called *homogeneous* if all constant terms are equal to zero, i.e., the homogeneous linear system is given by

$$\begin{aligned} \alpha_{11}x_1 + \alpha_{12}x_2 + \dots + \alpha_{1n}x_n &= 0, \\ \alpha_{21}x_1 + \alpha_{22}x_2 + \dots + \alpha_{2n}x_n &= 0, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \alpha_{m1}x_1 + \alpha_{m2}x_2 + \dots + \alpha_{mn}x_n &= 0. \end{aligned} \tag{5.58}$$

We shall consider some important *properties of homogeneous linear systems.*

(a) Any homogeneous system is consistent.

◄ The n-tuple $x_1 = 0$, $x_2 = 0$, $\dots$, $x_n = 0$ is the trivial solution of any homogeneous system. ►

(b) If the number n of the unknowns exceeds the number m of the equations the homogeneous system has nontrivial solutions.

◄ By definition the rank r of the system (5.58) satisfies the inequality $r \leqslant m < n$. Whence we infer that the linear system is indeterminate. ►

(c) The sum of solutions of the homogeneous system (5.58) is a solution of (5.58).

◄ Let $\gamma_1', \gamma_2', \dots, \gamma_n'$ and $\gamma_1'', \gamma_2'', \dots, \gamma_n''$ be two solutions of (5.58). This means that

$$\sum_{j=1}^{n} \alpha_{ij}\gamma_j' = 0 \quad \text{and} \quad \sum_{j=1}^{n} \alpha_{ij}\gamma_j'' = 0$$

for any i ($i = 1, 2, \dots, m$).

Since

$$\sum_{j=1}^{n} \alpha_{ij}(\gamma_j' + \gamma_j'') = \sum_{j=1}^{n} \alpha_{ij}\gamma_j' + \sum_{j=1}^{n} \alpha_{ij}\gamma_j'' = 0 \quad (i = 1, \dots, m),$$

the n-tuple

$$\gamma_1' + \gamma_1'', \quad \gamma_2' + \gamma_2'', \quad \dots, \quad \gamma_n' + \gamma_n''$$

i.e., the *sum of the solutions* of (5.58), is also the solution of the homogeneous linear system (5.58). ►

(d) The product of the solution of (5.58) by an arbitrary number is also a solution of (5.58).

◄ Let $\gamma_1, \gamma_2, \ldots, \gamma_n$ be a solution of (5.58), i.e.,

$$\sum_{j=1}^{n} \alpha_{ij}\gamma_j = 0 \quad (i = 1, 2, \ldots, m)$$

and let μ be an arbitrary number.

Then

$$\sum_{j=1}^{n} \alpha_{ij}(\mu\gamma_j) = \mu \sum_{j=1}^{n} \alpha_{ij}\gamma_j = 0 \quad (i = 1, 2, \ldots, m).$$

Whence we conclude that the n-tuple $\mu\gamma_1, \mu\gamma_2, \ldots, \mu\gamma_n$, i.e., the *product of the solution* $\gamma_1, \gamma_2, \ldots, \gamma_n$ *by the number* μ is the solution of the system (5.58). ►

In matrix form, the homogeneous linear system may be written as

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \hdotsfor{4} \\ \alpha_{m1} & \alpha_{m2} & & \alpha_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

or, as the matrix equation

$$\mathbf{AX} = \mathbf{0}. \tag{5.59}$$

Then the properties (c) and (d) may be formulated as follows.

(c$'$) If the column-vectors $\mathbf{\Gamma}'$ and $\mathbf{\Gamma}''$ are the solutions of the system (5.59), so is the column-vector $\mathbf{\Gamma}' + \mathbf{\Gamma}''$.

◄ Indeed, if $\mathbf{\Gamma}'$ and $\mathbf{\Gamma}''$ are the solutions of (5.59), then $\mathbf{A\Gamma}' = \mathbf{0}$, $\mathbf{A\Gamma}'' = \mathbf{0}$ and $\mathbf{A}(\mathbf{\Gamma}' + \mathbf{\Gamma}'') = \mathbf{A\Gamma}' + \mathbf{A\Gamma}'' = \mathbf{0} + \mathbf{0} = \mathbf{0}$, so that the vector-column $\mathbf{\Gamma}' + \mathbf{\Gamma}''$ is the solution of (5.59). ►

(d$'$) If the column-vector $\mathbf{\Gamma}$ is the solution of (5.59) so is $\mu\mathbf{\Gamma}$ where μ is an arbitrary number.

◄ Indeed, if $\mathbf{\Gamma}$ is the solution of (5.59) then $\mathbf{A\Gamma} = \mathbf{0}$ and $\mathbf{A}(\mu\mathbf{\Gamma}) = \mu(\mathbf{A\Gamma}) = \mu\mathbf{0} = \mathbf{0}$, so that the product of the column-vector $\mathbf{\Gamma}$ by an arbitrary number μ is the solution of (5.59). ►

The properties (c$'$) and (d$'$) imply that the solutions of the homogeneous linear system with the induced operations of addition and multiplication by an arbitrary number form a linear space.*) Let us acquaint ourselves with an important property of this linear space.

By the method of Gaussian elimination we may reduce the homogeneous linear system (5.58) to the form

---

*) The notion of linear space will be discussed in Chap. 6

$$\begin{aligned} x_1 + \beta_{12}x_2 + \ldots + \beta_{1r}x_r + \beta_{1,r+1}x_{r+1} + \ldots + \beta_{1n}x_n &= 0, \\ x_2 + \ldots + \beta_{2r}x_r + \beta_{2,r+1}x_{r+1} + \ldots + \beta_{2n}x_n &= 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ x_r + \beta_{r,r+1}x_{r+1} + \ldots + \beta_{rn}x_n &= 0. \end{aligned} \tag{5.60}$$

Allowing for renumbering of the unknowns we may consider $x_1, x_2, \ldots, x_r$ as the principal unknowns and the other unknowns as *indeterminate* or *disposal* unknowns.

Let the rank r of the system (5.60) be smaller than the number n of the unknowns, that is, $r < n$.

We compute $n - r$ basic solutions by giving to the indeterminate unknowns $x_{r+1}, x_{r+2}, \ldots, x_n$ the values from the following table

	x_{r+1}	x_{r+2}	...	x_{n+1}	x_n
1	1	0	...	0	0
2	0	1	...	0	0
...	...	...	...	...	...
$n-r-1$	0	0	...	1	0
$n-r$	0	0	...	0	1

(5.61)

Each row in (5.61) determines the solution of the system (5.60) so that we have the collection of $n - r$ solutions

$$\Gamma_1 = \begin{pmatrix} \gamma_{11} \\ \gamma_{21} \\ \vdots \\ \gamma_{r1} \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \Gamma_2 = \begin{pmatrix} \gamma_{12} \\ \gamma_{22} \\ \vdots \\ \gamma_{r2} \\ 0 \\ 1 \\ \vdots \\ 0 \\ 0 \end{pmatrix},$$

$$\ldots, \boldsymbol{\Gamma}_{n-r-1} = \begin{pmatrix} \gamma_{1,n-r-1} \\ \gamma_{2,n-r-1} \\ \vdots \\ \gamma_{r,n-r-1} \\ 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{pmatrix}, \quad \boldsymbol{\Gamma}_{n-r} = \begin{pmatrix} \gamma_{1,n-r} \\ \gamma_{2,n-r} \\ \vdots \\ \gamma_{r,n-r} \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

The solutions $\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \ldots, \boldsymbol{\Gamma}_{n-r}$ are linearly independent.

◀ Consider the linear combination

$$\mu_{r+1}\boldsymbol{\Gamma}_1 + \ldots + \mu_n\boldsymbol{\Gamma}_{n-r}$$

$$= \begin{pmatrix} \mu_{r+1}\gamma_{11} + \mu_{r+2}\gamma_{12} + \ldots + \mu_n\gamma_{1,n-r} \\ \mu_{r+1}\gamma_{21} + \mu_{r+2}\gamma_{22} + \ldots + \mu_n\gamma_{2,n-r} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \mu_{r+1}\gamma_{r1} + \mu_{r+2}\gamma_{r2} + \ldots + \mu_n\gamma_{r,n-r} \\ \mu_{r+1} \\ \mu_{r+2} \\ \ldots \\ \mu_{n-1} \\ \mu_n \end{pmatrix}. \tag{5.62}$$

It is easy to see that the linear combination (5.62) is equal to the zero column-vector if and only if $\mu_{r+1} = \mu_{r+2} = \ldots = \mu_{n-1} = \mu_n = 0$. Hence, it follows that the zero (trivial) solution of the system (5.60) is equal only to the trivial linear combination of the solutions $\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \ldots, \boldsymbol{\Gamma}_{n-r}$. ▶

The properties (c′) and (d′) imply that given arbitrary scalars μ_{r+1}, $\mu_{r+2}, \ldots, \mu_n$, the linear combination (5.62) is the solution of the system (5.60).

◀ Let

$$\boldsymbol{\Gamma} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_r \\ \mu_{r+1} \\ \vdots \\ \mu_n \end{pmatrix} \tag{5.63}$$

be an arbitrary solution of (5.60). Then (5.63) may be represented as a linear combination of the form (5.62).

Multiplying the solutions $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \ldots, \mathbf{\Gamma}_{n-r}$ by $\mu_{r+1}, \mu_{r+2}, \ldots, \mu_n$, respectively, and adding the multiples obtained, we get the solution of (5.60) as the linear combination (5.62).

Comparing (5.62) and (5.63), we easily see that both comprise the same values of the indeterminate unknowns $\mu_{r+1}, \mu_{r+2}, \ldots, \mu_n$. Observe that the values of the indeterminate unknowns uniquely determine the values of the principal unknowns. Hence, the solutions (5.62) and (5.63) are identical and

$$\mathbf{\Gamma} = \mu_{r+1}\mathbf{\Gamma}_1 + \mu_{r+2}\mathbf{\Gamma}_2 + \ldots + \mu_n\mathbf{\Gamma}_{n-r}. \quad ► \tag{5.64}$$

Therefore the solutions $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \ldots, \mathbf{\Gamma}_{n-r}$ of the homogeneous system (5.58) are such that (a) they are linearly independent and (b) any solution of (5.58) may be represented as a linear combination of $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \ldots, \mathbf{\Gamma}_{n-r}$.

Definition. Any collection of $n - r$ solutions of the homogeneous system (5.58) that satisfy the conditions (a) and (b) given above is called the *fundamental system of solutions* of (5.58).

Example. Compute the solutions of the system

$$\begin{cases} 3x_1 - 2x_2 + 2x_3 + 4x_4 = 0, \\ 6x_1 - 4x_2 + 4x_3 + 3x_4 = 0, \\ 9x_1 - 6x_2 + 3x_3 + 2x_4 = 0. \end{cases}$$

◄ By the method of Gaussian elimination we have

$$\begin{cases} 3x_1 - 2x_2 + 5x_3 + 4x_4 = 0, \\ \qquad\qquad - 6x_3 - 5x_4 = 0. \end{cases}$$

Let us choose x_2 and x_4 as indeterminate unknowns and arrange the table

x_1	x_2	x_3	x_4
$\frac{2}{3}$	1	0	0
$\frac{1}{18}$	0	$-\frac{5}{6}$	1

The fundamental system comprises the solutions

$$\mathbf{\Gamma}_1 = \begin{pmatrix} 2 \\ 3 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{\Gamma}_2 = \begin{pmatrix} 1 \\ 0 \\ -15 \\ 18 \end{pmatrix}.$$

Any solution of the given linear system can be expressed as

$$\mathbf{\Gamma} = \mu \begin{pmatrix} 2 \\ 3 \\ 0 \\ 0 \end{pmatrix} + \nu \begin{pmatrix} 1 \\ 0 \\ -15 \\ 18 \end{pmatrix},$$

where μ and ν are arbitrary numbers. ►

Summing up, it is important to mention that it suffices to find the fundamental system of solutions to describe the set of solutions of the homogeneous linear system since the set of all linear combinations of the fundamental solutions forms the desired set. It is also easy to see that table (5.61) represents the identity matrix of order $n - r$. However, this is one of the many ways to identify values of the indeterminate unknowns. Table (5.61) may contain *any* other nonsingular matrix of order $n - r$. Finally, we emphasize that any homogeneous linear system having nontrivial solutions possesses the fundamental system of solutions.

Methods of solving linear systems. In general the methods used to compute the solutions of linear systems can be classified as *direct* and *indirect* (or *iterative*).

Direct methods are based upon Gaussian elimination with some inessential modifications aimed at diminishing the number of operations required so that it remains proportional to n^3. When input data, i.e., the coefficient matrix **A** and the constant terms in (5.52), are defined exactly and the computations are free from rounding-off errors, the direct method yields the true solution as a result of finite sequence of steps or operations.

Indirect methods provide the approximate solutions by giving sequences of approximate solutions that start from some initial approximation. The complete cycle of computing the new approximate solution from the original one is called the *iteration*. So the indirect methods are often called the *iterative* ones. They involve a finite number of iterations to reach the approximate solutions with given precision.

In what follows we consider two methods, one direct and the other indirect, which are of practical interest in a variety of applications.

Gaussian elimination for linear systems with tridiagonal matrices. Of practical interest are linear systems with *tridiagonal* coefficient matrices which are square matrices $\mathbf{A} = (\alpha_{ij})$ whose nonzero elements are adjacent to their principal diagonals so that $\alpha_{ij} = 0$ if $|i - j| > 1$.

Let

$$\mathbf{A} = \begin{pmatrix} b_1 & c_1 & 0 & 0 \ldots & 0 & 0 & 0 \\ a_2 & b_2 & c_2 & 0 \ldots & 0 & 0 & 0 \\ 0 & a_3 & b_3 & c_3 \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 \ldots & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & 0 & 0 & 0 \ldots & 0 & a_n & b_n \end{pmatrix}$$

be a tridiagonal coefficient matrix. Then the corresponding linear system takes the form

$$
\begin{array}{l}
b_1 x_1 + c_1 x_2 \dotfill = d_1, \\
a_2 x_1 + b_2 x_2 + c_2 x_3 \dotfill = d_2, \\
\\
\dotfill \\
\qquad\qquad a_i x_{i-1} + b_i x_i + c_i x_{i+1} \dotfill = d_i, \\
\qquad\qquad \dotfill \\
\qquad\qquad\qquad\qquad a_n x_{n-1} + b_n x_n = d,
\end{array}
$$

where $i = 2, 3, \ldots, n - 1$, or

$$a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i, \tag{5.65}$$

where $i = 1, 2, \ldots, n$, $a_1 = 0$, $c_n = 0$.

Let us apply the method of Gaussian elimination to (5.65).

We assume that the unknowns x_i are related as

$$x_i = \alpha_{i+1} x_{i+1} + \beta_{i+1} \quad (i = 1, 2, \ldots, n), \tag{5.66}$$

where α_{i+1} and β_{i+1} are unknown quantities.

Diminishing all the subscripts in (5.66) by unity and substituting the obtained relations into (5.65), we get

$$a_i(\alpha_i x_i + \beta_i) + b_i x_i + c_i x_{i+1} = d_i.$$

Whence

$$x_i = -\frac{c_i}{b_i + a_i \alpha_i} x_{i+1} - \frac{a_i \beta_i - d_i}{b_i + a_i \alpha_i}.$$

This expression is identical to (5.66) if

$$\alpha_{i+1} = -\frac{c_i}{b_i + a_i \alpha_i} \quad \text{and} \quad \beta_{i+1} = -\frac{a_i \beta_i + d_i}{b_i + a_i \alpha_i}.$$

To start the computations we also need to know the values of α_1, β_1, x_{n+1}. These are usually set equal to

$$\alpha_1 = 0, \quad \beta_1 = 0, \quad x_{n+1} = 0.$$

Collecting all the formulas together, we may describe the method as consisting of

(a) *elimination stage*

$$\alpha_{i+1} = -\frac{c_i}{b_i + a_i \alpha_i}, \quad i = 1, 2, \ldots, n; \quad \alpha_1 = 0,$$

$$\beta_{i+1} = -\frac{a_i \beta_i - d_i}{b_i + a_i \alpha_i}, \quad i = 1, 2, \ldots, n; \quad \beta_1 = 0,$$

(b) *back-substitution stage*

$$x_i = \alpha_{i+1}x_{i+1} + \beta_{i+1}, \quad i = n, n-1, \ldots, 1, \; x_{n+1} = 0.$$

Therefore on completing the elimination stage we obtain the values of the quantities α_{i+1} and β_{i+1} ($i = 1, 2, \ldots, n$) and, consequently, the value of the unknown x_n. Then substituting x_n into the expression relating x_n and x_{n-1}, we get the value of x_{n-1}. Substituting x_{n-1} into the relation which comprises x_{n-1} and x_{n+2}, we get the value of x_{n-2}. Continuing this process, we finally compute the values for all the unknowns $x_1, x_2, \ldots, x_n$.

Method of simple iteration. Suppose that we are given a linear system (5.52) and start the computations from some initial approximate solution Y_0.

Let Y_k and Y_{k+1} be two approximate solutions obtained by completing the kth and $(k+1)$th iterations, respectively. Suppose that Y_k and Y_{k+1} are related as

$$\frac{Y_{k+1} - Y_k}{\tau} + AY_k = b \quad (k = 0, 1, 2, \ldots) \tag{5.67}$$

or

$$Y_{k+1} = Y_k - \tau(AY_k - b), \tag{5.68}$$

where τ ($\tau > 0$) is a real number called the *stationary parameter.*

The procedure which employs the recurrence relations (5.67) and (5.68) to compute the approximate solution of the system (5.52) is called the *method of simple iteration.*

In terms of unknowns the relation (5.68) is written as

$$y_{k+1}^{(i)} = y_k^{(i)} - \tau\left(\sum_{j=1}^{n} \alpha_{ij}y_k^{(i)} - \beta_j\right) \quad (i = 1, 2, \ldots, n). \tag{5.69}$$

The precision of the method is given by the value of the error

$$Z_k = Y_k - X$$

i.e., the value of the difference between the solution of (5.68) and the true solution X of the system (5.52).

Let ε ($\varepsilon > 0$) be a given relative error used to estimate the accuracy of the approximate solution Y_k. The process of the computations terminates when the following condition is satisfied

$$|Z_k| = \sqrt{\sum_{i=1}^{n} |z_k^{(i)}|^2} \leqslant \varepsilon \sqrt{\sum_{i=1}^{n} |z_0^{(i)}|^2} = \varepsilon |Z_0|.$$

The iterative method is said to *converge* if

$$\lim_{k\to\infty} |Z_k| = 0.$$

Suppose that the coefficient matrix $\mathbf{A} = (\alpha_{ij})$ of (5.52) is symmetric and all the eigenvalues of **A** are positive, i.e., $0 < \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n$ (see Chapter 6).

Theorem 5.10. *Whatever the initial approximation Y_0 the method of simple iteration converges if the stationary parameter τ satisfies the condition*

$$\tau < \frac{2}{\lambda_n}. \tag{5.70}$$

Example. Given the initial approximation

$$Y_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and the relative error $\varepsilon = 0.3$, use the method of simple iteration to compute the solution of the linear system

$$\begin{cases} 2x_1 + x_2 = 1, \\ x_1 + 2x_2 = 2, \end{cases}$$

or, in matrix form,

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

◄ Let us compute the eigenvalues of the symmetric matrix

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

We have

$$\begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = (\lambda - 2)^2 - 1 = 0.$$

Whence $\lambda_1 = 1$ and $\lambda_2 = 3$.

Put

$$\tau = \frac{1}{2} = 0.5.$$

Thus condition (5.70) is satisfied. Then from (5.68) we have

$$Y_{k+1} = Y_k - 0.5\left\{\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} Y_k - \begin{pmatrix} 1 \\ 2 \end{pmatrix}\right\}.$$

Whence

$$Y_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + 0.5 \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 1 \end{pmatrix},$$

$$Y_2 = \begin{pmatrix} 0.5 \\ 1 \end{pmatrix} - 0.5 \left\{ \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 0.5 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right\}$$

$$= \begin{pmatrix} 0.5 \\ 1 \end{pmatrix} - 0.5 \begin{pmatrix} 2 - 1 \\ 2.5 - 2 \end{pmatrix} = \begin{pmatrix} 0.5 - 0.5 \\ 1 - 0.25 \end{pmatrix} = \begin{pmatrix} 0 \\ 0.75 \end{pmatrix}.$$

Let us estimate the precision of the approximate solutions. Since

$$X = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

is the true solution of the system, then

$$|Z_0| = |Y_0 - X| = \sqrt{1^2} = 1,$$
$$|Z_1| = |Y_1 - X| = \sqrt{(0.5)^2} = 0.5,$$
$$|Z_2| = |Y_2 - X| = \sqrt{(0.25)^2} = 0.25.$$

Whence we infer that the approximate solution

$$Y_2 = \begin{pmatrix} 0 \\ 0.75 \end{pmatrix}$$

obtained after the second iteration meets the accuracy required. ►

Exercises

1. Multiply the matrix **A** by the matrix **B**:

(a) $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$; (b) $\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$.

2. Multiply the matrix $\begin{pmatrix} 1 & a & c \\ 0 & 1 & b \\ 0 & 0 & 1 \end{pmatrix}$ by itself.

3. Compute the product **AB** and **BA** for $\mathbf{A} = (2 \quad -3 \quad 0)$ and $\mathbf{B} = \begin{pmatrix} 4 \\ 3 \\ 1 \end{pmatrix}$.

4. Reduce the matrices to schematic forms: (a) $\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 2 \\ 1 & 1 & 3 & 1 \\ 1 & 2 & 1 & 4 \end{pmatrix}$;

(b) $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 5 & 6 \\ 5 & 6 & 7 & 8 \\ 31 & 23 & 55 & 42 \end{pmatrix}$.

5. Compute the values of the determinants: (a) $\begin{vmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{vmatrix}$;

(b) $\begin{vmatrix} 1 & 2 & 3 & 4 \\ 3 & 6 & 8 & 11 \\ 7 & 13 & 20 & 26 \\ 31 & 23 & 55 & 42 \end{vmatrix}$; (c) $\begin{vmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 5 & 6 \\ 5 & 6 & 7 & 9 \\ 31 & 23 & 55 & 42 \end{vmatrix}$.

6. Invert the matrices: (a) $\begin{pmatrix} 5 & -4 \\ -8 & 6 \end{pmatrix}$, (b) $\begin{pmatrix} 1 & -3 & -1 \\ -2 & 7 & 2 \\ 3 & 2 & -4 \end{pmatrix}$,

(c) $\begin{pmatrix} 0 & 1 & 1 & 1 \\ -1 & 0 & 1 & 1 \\ -1 & -1 & 0 & 1 \\ -1 & -1 & -1 & 0 \end{pmatrix}$.

7. Find the ranks of: (a) $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix}$; (b) $\begin{pmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{pmatrix}$;

(c) $\begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ -1 & 0 & 0 & 1 \end{pmatrix}$.

8. Compute the solutions of the linear systems: (a) $\begin{cases} x_1 + x_2 - x_3 = 1, \\ x_1 - x_2 + x_3 = 1, \\ -x_1 + x_2 - x_3 = 1; \end{cases}$

(b) $\begin{cases} x_1 - x_2 + 2x_3 = 2, \\ x_1 + 2x_2 + x_3 = -2, \\ 2x_1 - x_2 - x_3 = -2; \end{cases}$ (c) $\begin{cases} 7x_1 - 5x_2 - 2x_3 - 4x_4 = 8, \\ -3x_1 + 2x_2 + x_3 + 2x_4 = -3, \\ 2x_1 - x_2 - x_3 - 2x_4 = 1, \\ -x_1 + x_3 + 2x_4 = 1; \end{cases}$

(d) $\begin{cases} -3x_1 + x_2 + x_3 = 0, \\ 5x_1 + x_2 - 2x_3 = 2, \\ -2x_1 - 2x_2 + x_3 = -3. \end{cases}$

9. Compute the solutions of the systems: (a) $\begin{cases} 3x_1 + 2x_2 + x_3 = 0, \\ 7x_1 + 6x_2 + 5x_3 = 0, \\ 5x_1 + 4x_2 + 3x_3 = 0; \end{cases}$

$$\text{(b)}\begin{cases} -x_1 \phantom{{}-x_2} + x_3 + 2x_4 = 0, \\ - x_2 + x_3 + 2x_4 = 0, \\ 7x_1 - 5x_2 - 2x_3 - 4x_4 = 0, \\ 4x_1 - 3x_2 - x_3 - 2x_4 = 0; \end{cases} \quad \text{(c)}\begin{cases} x_1 + 4x_2 + 2x_3 \phantom{{}+2x_4} - 3x_5 = 0, \\ 2x_1 + 9x_2 + 5x_3 + 2x_4 + x_5 = 0, \\ x_1 + 3x_2 + x_3 - 2x_4 - 9x_5 = 0. \end{cases}$$

Answers

1. (a) $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$; (b) $\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$. **2.** $\begin{pmatrix} 1 & 2a & ab + 2c \\ 0 & 1 & 2b \\ 0 & 0 & 1 \end{pmatrix}$.

3. $\mathbf{AB} = (-1)$, $\mathbf{BA} = \begin{pmatrix} 8 & -12 & 0 \\ 6 & -9 & 0 \\ 2 & -3 & 0 \end{pmatrix}$.

4. (a) $\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}$; (b) $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -2 & -4 & -6 \\ 0 & 0 & 40 & 35 \\ 0 & 0 & 0 & 0 \end{pmatrix}$.

5. (a) -12; (b) 5; (c) 80. **6.** (a) $-\dfrac{1}{2}\begin{pmatrix} 6 & 4 \\ 8 & 5 \end{pmatrix}$; (b) $\begin{pmatrix} 32 & 14 & -1 \\ 2 & 1 & 0 \\ 25 & 11 & -1 \end{pmatrix}$;

(c) $\begin{pmatrix} 0 & -1 & 1 & -1 \\ 1 & 0 & -1 & 1 \\ -1 & 1 & 0 & -1 \\ 1 & -1 & 1 & 0 \end{pmatrix}$.

7. (a) 2; (b) 2; (c) 3; **8.** (a) $x_1 = 1$, $x_2 = 1$, $x_3 = 1$; (b) $x_1 = -1$, $x_2 = -1$, $x_3 = 1$; (c) $x_1 = -1 - \gamma_3 + 2\gamma_4$, $x_2 = -3 + \gamma_3 + 2\gamma_4$, $x_3 = \gamma_3$, $x_4 = \gamma_4$; (d) the system is inconsistent. **9.** (a) $x_1 = x_3$, $x_2 = -2x_3$; (b) $x_1 = x_2 = x_3 + 2x_4$; or

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \lambda \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix};$$

or $$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \lambda \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix} + \mu \begin{pmatrix} 2 \\ 2 \\ 0 \\ 1 \end{pmatrix};$$

(c) $$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \lambda \begin{pmatrix} 2 \\ -1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \mu \begin{pmatrix} 8 \\ -2 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

Chapter 6

Linear Spaces and Linear Operators

6.1. The Concept of Linear Space

A set **V** of mathematical objects **x**, **y**, **z**, ..., called the elements of **V**, is said to form a *linear real* or *complex space* if (1) given two arbitrary elements **x** and **y** in **V** there exists an element **x** + **y** in **V** called the *sum* of **x** and **y**, and (2) given an arbitrary element **x** in **V** and an arbitrary real or complex α there exists an element $\alpha\mathbf{x}$ in **V** called the *multiple* of **x** by α, and the induced operations of addition and multiplication by a scalar satisfy the following eight conditions

(a) $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$;

(b) $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$;

(c) there exists an element $\boldsymbol{\theta}$ in **V** such that for any **x** in **V** there holds $\mathbf{x} + \boldsymbol{\theta} = \mathbf{x}$;

(d) for any **x** in **V** there exists an element $(-\mathbf{x})$ in **V** such that $\mathbf{x} + (-\mathbf{x}) = \boldsymbol{\theta}$;

(e) $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$;

(f) $(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$;

(g) $\alpha(\beta\mathbf{x}) = (\alpha\beta)\mathbf{x}$;

(h) $1 \cdot \mathbf{x} = \mathbf{x}$.

The elements **x**, **y**, **z**, ... of the linear space **V** are often called the *vectors*, the element $\boldsymbol{\theta}$ is called the *zero vector* and the element $(-\mathbf{x})$ is called the *additive inverse* of the vector **x**. The linear space is also called the *vector space.*

Examples of linear spaces. The following sets of mathematical objects form linear spaces:

(1) The set $\mathbf{V}_3$ of geometric free vectors with the induced operations of vector addition and multiplication of a vector by a scalar described in Chap. 2 (Fig. 6.1) as well as the set $\mathbf{V}_1$ of vectors in a line and the set $\mathbf{V}_2$ of vectors in a plane.

(2) The set of sequences $(\xi^1, \xi^2, ..., \xi^n)$ of n real numbers. The operations of addition and multiplication by a real number are induced as

(a) *Addition*

$$(\xi^1, \xi^2, ..., \xi^n) + (\eta^1, \eta^2, ..., \eta^n)$$
$$= (\xi^1 + \eta^1, \xi^2 + \eta^2, ..., \xi^n + \eta^n);$$

(b) *Multiplication by a real number*

$$\lambda(\xi^1, \xi^2, \ldots, \xi^n) = (\lambda\xi^1, \lambda\xi^2, \ldots, \lambda\xi^n).$$

In symbols $\mathbb{R}^n$ designates the *n-dimensional real coordinate space* (Fig. 6.2).

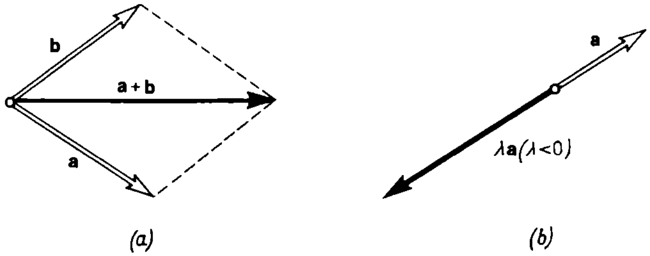


Fig. 6.1

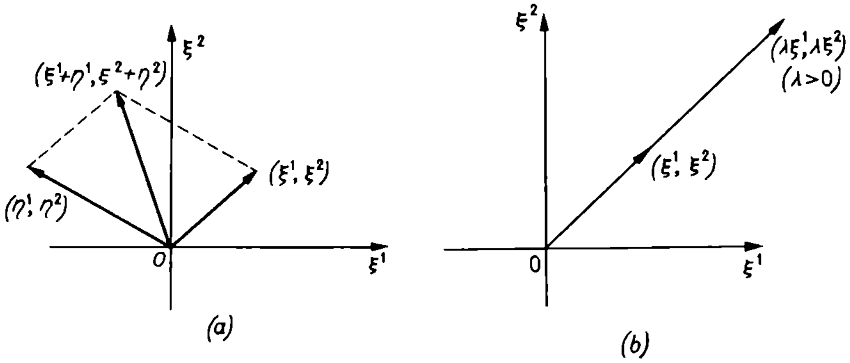


Fig. 6.2

(3) The set $\mathbb{R}^{m\times n}$ of $m \times n$ matrices, discussed in Chap. 5, with the operations of matrix addition

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \multicolumn{4}{c}{\dotfill} \\ \alpha_{m1} & \alpha_{m2} & \ldots & \alpha_{mn} \end{pmatrix} + \begin{pmatrix} \beta_{11} & \beta_{12} & \ldots & \beta_{1n} \\ \beta_{21} & \beta_{22} & \ldots & \beta_{2n} \\ \multicolumn{4}{c}{\dotfill} \\ \beta_{m1} & \beta_{m2} & \ldots & \beta_{mn} \end{pmatrix}$$

$$= \begin{pmatrix} \alpha_{11}+\beta_{11} & \alpha_{12}+\beta_{12} & \ldots & \alpha_{1n}+\beta_{1n} \\ \alpha_{21}+\beta_{21} & \alpha_{22}+\beta_{22} & \ldots & \alpha_{2n}+\beta_{2n} \\ \multicolumn{4}{c}{\dotfill} \\ \alpha_{m1}+\beta_{m1} & \alpha_{m2}+\beta_{m2} & \ldots & \alpha_{mn}+\beta_{mn} \end{pmatrix}$$

and multiplication of a matrix by a real number

$$\lambda\begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \dots & \dots & \dots & \dots \\ \alpha_{m1} & \alpha_{m2} & \dots & \alpha_{mn} \end{pmatrix} = \begin{pmatrix} \lambda\alpha_{11} & \lambda\alpha_{12} & \dots & \lambda\alpha_{1n} \\ \lambda\alpha_{21} & \lambda\alpha_{22} & \dots & \lambda\alpha_{2n} \\ \dots & \dots & \dots & \dots \\ \lambda\alpha_{m1} & \lambda\alpha_{m2} & \dots & \lambda\alpha_{mn} \end{pmatrix}.$$

In particular, the set $\mathbb{R}^{1 \times n}$ of row-vectors, i.e., $1 \times n$ matrices, and the set $\mathbb{R}^{m \times 1}$ of column-vectors, i.e., $m \times 1$ matrices, form linear spaces.

(4) The set $\mathbf{C}(0, 1)$ of real functions continuous on the interval $(0, 1)$ with the usual operations of addition of functions and multiplication of a function by a real number.

In all these cases the validity of the conditions (a)-(h) are directly verified.

Properties of linear spaces. (a) The zero vector $\boldsymbol{\theta}$ is uniquely defined. ◄ Let $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ be zero vectors in $\mathbf{V}$ and let $\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2$ be a sum of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Since $\boldsymbol{\theta}_2$ is the zero vector the condition (c) implies that $\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2 = \boldsymbol{\theta}_1$. The vector $\boldsymbol{\theta}_1$ is also zero. Then $\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2 = \boldsymbol{\theta}_2 + \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. ►

(b) For any vector $\mathbf{x}$ the additive inverse vector $(-\mathbf{x})$ of $\mathbf{x}$ is uniquely defined.

◄ Let $\mathbf{x}^-$ and $\mathbf{x}_-$ be two additive vectors inverse to $\mathbf{x}$. Let us show that $\mathbf{x}^-$ and $\mathbf{x}_-$ are equal.

Consider the sum $\mathbf{x}_- + \mathbf{x} + \mathbf{x}^-$. Since $\mathbf{x}^-$ is inverse to the vector $\mathbf{x}$ the condition (a) implies that

$$\mathbf{x}_- + \mathbf{x} + \mathbf{x}^- = \mathbf{x}_- + (\mathbf{x} + \mathbf{x}^-) = \mathbf{x}_- + \boldsymbol{\theta} = \mathbf{x}_-.$$

Analogously, we have

$$\mathbf{x}_- + \mathbf{x} + \mathbf{x}^- = (\mathbf{x}_- + \mathbf{x}) + \mathbf{x}^- = \boldsymbol{\theta} + \mathbf{x}^- = \mathbf{x}^-. \text{ ►}$$

It is easy to verify the following properties:

(c) for any vector $\mathbf{x}$ there holds $0\mathbf{x} = \boldsymbol{\theta}$;

(d) for any vector $\mathbf{x}$ there holds $-\mathbf{x} = (-1)\mathbf{x}$;

(e) for any real number α there holds $\alpha\boldsymbol{\theta} = \boldsymbol{\theta}$;

(f) the identity $\alpha\mathbf{x} = \boldsymbol{\theta}$ implies that either $\alpha = 0$ or $\mathbf{x} = \boldsymbol{\theta}$.

6.2 Linear Subspaces

A non-empty subset $\mathbf{W}$ of a linear space $\mathbf{V}$ is said to form a *linear subspace* in $\mathbf{V}$ if given arbitrary vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbf{W}$ and any number α, the following conditions are satisfied:

(i) $\mathbf{x} + \mathbf{y}$ is an element of $\mathbf{W}$ and (ii) $\alpha\mathbf{x}$ is an element of $\mathbf{W}$. The set $\mathbf{W}$ is said to be *closed* with respect to the operations (i) and (ii).

Examples of linear subspaces. (1) The set $\mathbf{V}_2$ of vectors in a plane forms a subspace of the linear space $\mathbf{V}_3$.

(2) The solution of the homogeneous system of m linear equations in n unknowns

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \dots & \dots & \dots & \dots \\ \alpha_{m1} & \alpha_{m2} & \dots & \alpha_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

or, in matrix form,

$$\mathbf{AX} = \mathbf{0} \qquad (*)$$

forms a linear subspace.

◄ In Chap. 5 we have shown that the sum of the solutions of (*) and the multiple of the solutions of (*) by a number are also the solutions of (*). Whence we infer that the set of the solutions of (*) forms a linear subspace of the linear space $\mathbb{R}^{n \times 1}$. ►

(3) The set of all functions $f(x)$ such that each $f(x)$ is equal to zero at $x = 0$, i.e., $f(0) = 0$, forms a subspace in the linear space $\mathbf{C}(-1, 1)$ of real-valued functions continuous on the interval $(-1, 1)$.

◄ It is easy to verify that the set

$$\mathbf{C}_0(-1, 1) = \{f(x) | f(0) = 0\}$$

is a linear subspace since both the sum $f(x) + g(x)$ and the multiple $\alpha f(x)$ are equal to zero at $x = 0$ when $f(0) = g(0) = 0$. ►

From the definition of linear subspace it follows that if $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_q$ are vectors in $\mathbf{W}$ so is any linear combination $\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \dots + \alpha_q\mathbf{x}_q$. Also, for any linear subspace $\mathbf{W}$ the conditions (a)-(h) are satisfied. This means that *any linear subspace of linear space is itself a linear space.* It suffices to show that both the zero vector and the additive vector inverse to an arbitrary vector in $\mathbf{W}$ belong to $\mathbf{W}$.

◄ Let $\mathbf{x}$ be an arbitrary vector in $\mathbf{W}$. Then multiplying $\mathbf{x}$ by 0 and -1, we get the zero vector $\boldsymbol{\theta} = 0\mathbf{x}$ and the additive vector $-\mathbf{x} = (-1)\mathbf{x}$ inverse to $\mathbf{x}$. Obviously, the other conditions are also satisfied. ►

Sums and intersections of linear subspaces. Let $\mathbf{V}$ be a linear space and let $\mathbf{W}_1$ and $\mathbf{W}_2$ be linear subspaces in $\mathbf{V}$.

The *sum* $\mathbf{W}_1 + \mathbf{W}_2$ of the linear subspaces $\mathbf{W}_1$ and $\mathbf{W}_2$ is the set of all vectors $\mathbf{x}$ in $\mathbf{V}$ such that

$$\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2, \qquad (6.1)$$

where $\mathbf{x}_1$ is a vector in $\mathbf{W}_1$ and $\mathbf{x}_2$ is a vector in $\mathbf{W}_2$.

In symbols, we write the sum of $\mathbf{W}_1$ and $\mathbf{W}_2$ as

$$\mathbf{W}_1 + \mathbf{W}_2 = \{\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2 | \mathbf{x}_1 \in \mathbf{W}_1, \ \mathbf{x}_2 \in \mathbf{W}_2\}.$$

We show that the sum $\mathbf{W}_1 + \mathbf{W}_2$ is a linear subspace in $\mathbf{V}$.

◄ Let $\mathbf{x}$ and $\mathbf{y}$ be two arbitrary vectors in $\mathbf{W}_1 + \mathbf{W}_2$. By definition of the sum of linear spaces there exist $\mathbf{x}_1$ and $\mathbf{y}_1$ in $\mathbf{W}_1$, and $\mathbf{x}_2$ and $\mathbf{y}_2$ in $\mathbf{W}_2$ such that

$$\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2 \quad \text{and} \quad \mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2.$$

Whence the sum $\mathbf{x} + \mathbf{y}$ may be written as

$$\mathbf{x} + \mathbf{y} = (\mathbf{x}_1 + \mathbf{x}_2) + (\mathbf{y}_1 + \mathbf{y}_2) = (\mathbf{x}_1 + \mathbf{y}_1) + (\mathbf{x}_2 + \mathbf{y}_2).$$

Since $\mathbf{x}_1 + \mathbf{y}_1$ belongs to $\mathbf{W}_1$ and $\mathbf{x}_2 + \mathbf{y}_2$ belongs to $\mathbf{W}_2$ the sum $\mathbf{x} + \mathbf{y}$ is a vector in $\mathbf{W}_1 + \mathbf{W}_2$.

Similarly, we can prove that $\alpha\mathbf{x}$ is a vector in $\mathbf{W}_1 + \mathbf{W}_2$. ►

The sum of the linear spaces $\mathbf{W}_1$ and $\mathbf{W}_2$ is said to be *direct* if the expansion (6.1) is unique for any vector $\mathbf{x}$ in $\mathbf{W}_1 + \mathbf{W}_2$ (Fig. 6.3).

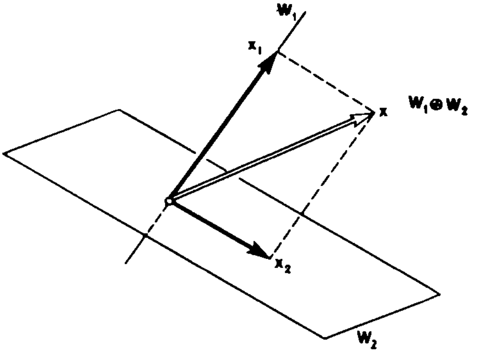


Fig. 6.3

In symbols, we write $\mathbf{W}_1 \oplus \mathbf{W}_2$ to designate the direct sum of $\mathbf{W}_1$ and $\mathbf{W}_2$.

The *intersection* $\mathbf{W}_1 \cap \mathbf{W}_2$ of the linear subspaces $\mathbf{W}_1$ and $\mathbf{W}_2$ of the linear space $\mathbf{V}$ is the set of all vectors $\mathbf{x}$ that belong both to $\mathbf{W}_1$ and to $\mathbf{W}_2$.

Properties of sums and intersections of linear subspaces. (a) The intersection $\mathbf{W}_1 \cap \mathbf{W}_2$ is non-empty since $\mathbf{W}_1 \cap \mathbf{W}_2$ always contains the zero vector.

(b) The intersection $\mathbf{W}_1 \cap \mathbf{W}_2$ is a linear subspace of the linear space $\mathbf{V}$.

(c) If the zero vector is the only vector contained both in $\mathbf{W}_1$ and in $\mathbf{W}_2$, the sum of $\mathbf{W}_1$ and $\mathbf{W}_2$ is direct (denoted by $\mathbf{W}_1 \oplus \mathbf{W}_2$).

Linear span. Let $\mathbf{V}$ be a linear space and $\mathbf{X}$ be a subset of $\mathbf{V}$. The set

$$\mathbf{L}(\mathbf{X}) = \left\{\mathbf{y} = \sum_{j=1}^{q} \alpha_j\mathbf{x}_j \middle| \mathbf{x}_j \in \mathbf{X}; \quad \alpha_j \in \mathbb{R}; \quad q = 1, 2, \ldots\right\}$$

of all linear combinations of vectros $\mathbf{y}$ in $\mathbf{X}$ is called the *linear span* $\mathbf{L}(\mathbf{X})$ of $\mathbf{X}$. In other words, the linear span $\mathbf{L}(\mathbf{X})$ of $\mathbf{X}$ is the set of all vectors $\mathbf{y}$ that may be represented as linear combinations of vectors in $\mathbf{X}$.

Properties of linear spans. (a) The linear span $\mathbf{L}(\mathbf{X})$ contains the set $\mathbf{X}$.

(b) The linear span $\mathbf{L}(\mathbf{X})$ forms a linear subspace in $\mathbf{V}$.

◄ This follows from the fact that the sum of linear combinations of vectors in $\mathbf{X}$ and the multiples of linear combinations by an arbitrary number are linear combinations of vectors in $\mathbf{X}$. ►

(c) The linear span $\mathbf{L}(\mathbf{X})$ is the smallest linear subspace containing the set $\mathbf{X}$.

In other words, if the linear subspace $\mathbf{W}$ contains the set $\mathbf{X}$, then $\mathbf{W}$ also contains the linear span $\mathbf{L}(\mathbf{X})$ of $\mathbf{X}$.

◄ Indeed, an arbitrary linear combination $\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \dots + \alpha_q\mathbf{x}_q$ of vectors in $\mathbf{X}$, being an element of $\mathbf{L}(\mathbf{X})$, is also contained in $\mathbf{W}$. ►

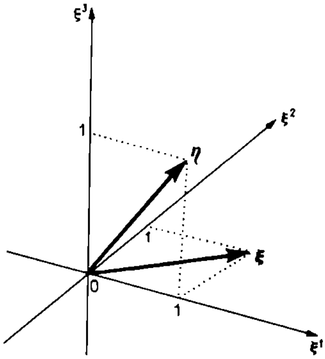


Fig. 6.4

Examples. (1) Let $\boldsymbol{\xi} = (1, 1, 0)$ and $\boldsymbol{\eta} = (1, 0, 1)$ be two vectors in the linear space $\mathbb{R}^3$. Then the set of solutions of the equation

$$\xi^1 + \xi^2 - \xi^3 = 0 \tag{6.2}$$

is the linear span $\mathbf{L}(\boldsymbol{\xi}, \boldsymbol{\eta})$ of the vectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$.

◄ Indeed, the triples (1, 1, 0) and (1, 0, 1) form the fundamental system of solutions of the homogeneous equation (6.2). Hence each solution of (6.2) is a linear combination of fundamental solutions (Fig. 6.4). ►

(2) Let $\mathbf{C}(-\infty, \infty)$ be a linear space of real-valued functions continuous

at every point of the number line and let

$$\mathbf{X} = \{1, x, ..., x^n\}$$

be a set of monomials 1, x, ..., x^n.

Then the set of polynomials of degree not exceeding n with real coefficients forms the linear span $\mathbf{L}(\mathbf{X})$ of $\mathbf{X}$.

In symbols, $\mathbf{L}(1, x, ..., x^n)$ designates the linear span of the set of monomials 1, x, ..., x^n and $\mathbf{M}_n$ stands for the set of polynomials of degree not exceeding n with real coefficients.

6.3 Linearly Dependent Vectors

Definition. The vectors $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_q$ in the linear space $\mathbf{V}$ are called *linearly dependent* if there exist numbers $\alpha_1, \alpha_2, ..., \alpha_q$, not all zero, such that

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + ... + \alpha_q\mathbf{x}_q = \boldsymbol{\theta}. \tag{6.3}$$

If the equality (6.3) is satisfied by $\alpha_1 = \alpha_2 = ... = \alpha_q = 0$ only, the vectors $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_q$ are called *linearly independent.*

The following propositions are valid.

Theorem 6.1. *The vectors* $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_q$ ($q \geqslant 2$) *are linearly dependent if and only if at least one of these vectors may be represented as a linear combination of the others.*

◄ Assume that the vectors $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_q$ are linearly dependent. For definiteness we set $\alpha_q \neq 0$ in (6.3). Transposing to the right all the terms in (6.3) except the qth one and dividing all the terms by $\alpha_q \neq 0$, we have

$$\mathbf{x}_q = -\frac{\alpha_1}{\alpha_q}\mathbf{x}_1 - \frac{\alpha_2}{\alpha_q}\mathbf{x}_2 - ... - \frac{\alpha_{q-1}}{\alpha_q}\mathbf{x}_{q-1},$$

i.e., the vector $\mathbf{x}_q$ is a linear combination of the vectors $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_q$.

Conversely, if one of the vectors is equal to the linear combination of the others

$$\beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + ... + \beta_{q-1}\mathbf{x}_{q-1} = \mathbf{x}_q$$

then transposing $\mathbf{x}_q$ to the left, we have the linear combination

$$\beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + ... + \beta_{q-1}\mathbf{x}_{q-1} + (-1)\mathbf{x}_q = \boldsymbol{\theta}$$

that contains at least one nonzero coefficient ($-1 \neq 0$). This means that the vectors $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_q$ are linearly dependent. ►

Theorem 6.2. *Let* $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_q$ *be linearly independent vectors and let* $\mathbf{y} = \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + ... + \alpha_q\mathbf{x}_q$. *Then the coefficients* $\alpha_1, \alpha_2, ..., \alpha_q$ *are uniquely defined by* $\mathbf{y}$.

◄ Let

$$\mathbf{y} = \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \dots + \beta_q\mathbf{x}_q.$$

Then

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \dots + \alpha_q\mathbf{x}_q = \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \dots + \beta_q\mathbf{x}_q.$$

Whence

$$(\alpha_1 - \beta_1)\mathbf{x}_1 + (\alpha_2 - \beta_2)\mathbf{x}_2 + \dots + (\alpha_q - \beta_q)\mathbf{x}_q = \boldsymbol{\theta}.$$

Since the vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_q$ are linearly independent, $\alpha_1 - \beta_1 = \alpha_2 - \beta_2 = \dots = \alpha_q - \beta_q = 0$. Hence, $\alpha_1 = \beta_1$, $\alpha_2 = \beta_2$, ..., $\alpha_q = \beta_q$. ►

Theorem 6.3. *A collection of vectors containing linearly dependent vectors is linearly dependent.*

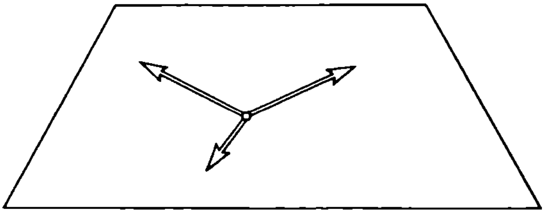


Fig. 6.5

◄ Let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_q, \mathbf{x}_{q+1}, \dots, \mathbf{x}_m$ be a collection of vectors such that the first q vectors are linearly dependent. Then

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \dots + \alpha_q\mathbf{x}_q = \boldsymbol{\theta}$$

and not all $\alpha_1, \alpha_2, \dots, \alpha_q$ are equal to zero. Adding to this linear combination the multiples of $\mathbf{x}_{q+1}, \mathbf{x}_{q+2}, \dots, \mathbf{x}_m$ by zero, we obtain the linear combination

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \dots + \alpha_q\mathbf{x}_q + 0\mathbf{x}_{q+1} + \dots + 0\mathbf{x}_m = \boldsymbol{\theta},$$

where not all α_i are equal to zero. ►

Example. The vectors in $\mathbf{V}_3$ are linearly dependent if and only if they are coplanar (Fig. 6.5).

6.4 Basis and Dimension

The sequence of vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ in a linear space $\mathbf{V}$ is called the ***basis*** of $\mathbf{V}$ if $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ are linearly independent and any vector in $\mathbf{V}$ may be expressed as a linear combination of $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$.

We speak of the sequence of $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ to mean an ordered n-tuple of vectors. It is easy to see that by permutation of n vectors we obtain $n!$ ordered sequences (collections) of n vectors.

Example. Let **a**, **b**, **c** be three noncoplanar vectors in $\mathbf{V}_3$ (Fig. 6.6). Then the ordered triples (**a**, **b**, **c**), (**b**, **c**, **a**), (**c**, **a**, **b**), (**b**, **a**, **c**), (**a**, **c**, **b**) and (**c**, **b**, **a**) form distinct bases in $\mathbf{V}_3$.

Let $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ be a basis in **V**. Then given an arbitrary vector **x** in **V**, there exists a collection of n numbers $\xi^1, \xi^2, \ldots, \xi^n$ such that

$$\mathbf{x} = \xi^1\mathbf{e}_1 + \xi^2\mathbf{e}_2 + \ldots + \xi^n\mathbf{e}_n = \sum_{i=1}^{n} \xi^i\mathbf{e}_i.$$

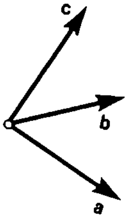


Fig. 6.6

By Theorem 6.2 the numbers $\xi^1, \xi^2, \ldots, \xi^n$, called the *coordinates* of the vector **x** relative to the basis **e**, are uniquely defined.

Let $\mathbf{x} = \sum_{i=1}^{n} \xi^i\mathbf{e}_i$ and $\mathbf{y} = \sum_{i=1}^{n} \eta^i\mathbf{e}_i$ be vectors in **V**. Then

$$\mathbf{x} + \mathbf{y} = \sum_{i=1}^{n} \xi^i\mathbf{e}_i + \sum_{i=1}^{n} \eta^i\mathbf{e}_i = \sum_{i=1}^{n} (\xi^i + \eta^i)\mathbf{e}_i$$

and, for any number α,

$$\alpha\mathbf{x} = \alpha \sum_{i=1}^{n} \xi^i\mathbf{e}_i = \sum_{i=1}^{n} (\alpha\xi^i)\mathbf{e}_i.$$

Thus, addition of vectors becomes addition of their respective coordinates and multiplication of a vector by a scalar becomes multiplication of the coordinates by this scalar.

The coordinates of a vector may conveniently be written in the form of a column-vector. For example,

$$\begin{pmatrix} \xi^1 \\ \xi^2 \\ \vdots \\ \xi^n \end{pmatrix}$$

is the column-vector of the coordinates of the vector $\mathbf{x} = \sum_{i=1}^{n} \xi^i \mathbf{e}_i$ relative to the basis $\mathbf{e}$.

Let

$$\mathbf{x}_1 = \sum_{i=1}^{n} \xi_1^i \mathbf{e}_i, \quad \mathbf{x}_2 = \sum_{i=1}^{n} \xi_2^i \mathbf{e}_i, \ \ldots, \ \mathbf{x}_q = \sum_{i=1}^{n} \xi_q^i \mathbf{e}_i$$

be expansions of the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q$ relative to the basis $\mathbf{e}$ and let

$$\begin{pmatrix} \xi_1^1 \\ \xi_1^2 \\ \vdots \\ \xi_1^n \end{pmatrix} \begin{pmatrix} \xi_2^1 \\ \xi_2^2 \\ \vdots \\ \xi_2^n \end{pmatrix}, \ldots, \begin{pmatrix} \xi_q^1 \\ \xi_q^2 \\ \vdots \\ \xi_q^n \end{pmatrix}$$

be column-vectors of the coordinates of $\mathbf{x}_1, \ldots, \mathbf{x}_2, \ldots, \mathbf{x}_q$ relative to $\mathbf{e}$.

Theorem 6.4. *The vectors* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q$ *are linearly dependent if and only if so are the column-vectors of their coordinates relative to some basis* $\mathbf{e}$.

◄ Assume that the linear combination $\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2 + \ldots + \lambda_q\mathbf{x}_q$ of the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q$ is equal to the zero vector, i.e.,

$$\sum_{k=1}^{q} \lambda_k \mathbf{x}_k = \boldsymbol{\theta}$$

or, in explicit form,

$$\sum_{k=1}^{q} \lambda_k \left(\sum_{i=1}^{n} \xi_k^i \mathbf{e}_i \right) = \sum_{i=1}^{n} \left(\sum_{k=1}^{q} \lambda_k \xi_k^i \right) \mathbf{e}_i = \theta. \tag{6.4}$$

Recall that the expansion of a vector relative to a given basis is unique. Then from (6.4) we have

$$\sum_{k=1}^{q} \lambda_k \xi_k^1 = 0, \quad \sum_{k=1}^{q} \lambda_k \xi_k^2 = 0, \ \ldots, \ \sum_{k=1}^{q} \lambda_k \xi_k^n = 0$$

or, equivalently,

$$\lambda_1 \begin{pmatrix} \xi_1^1 \\ \xi_1^2 \\ \vdots \\ \xi_1^n \end{pmatrix} + \lambda_2 \begin{pmatrix} \xi_2^1 \\ \xi_2^2 \\ \vdots \\ \xi_2^n \end{pmatrix} + \ldots + \lambda_q \begin{pmatrix} \xi_q^1 \\ \xi_q^2 \\ \vdots \\ \xi_q^n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{6.5}$$

This means that the linear combination of column-vectors of coordinates of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q$ is equal to the zero column-vector.

If we suppose that (6.5) holds, then reversing our arguments, we arrive at formula (6.4). Whence we infer that if a nontrivial linear combination of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q$ (i.e. a linear combination with $\lambda_1, \lambda_2, \ldots, \lambda_q$ not all equal to zero) vanishes, then a nontrivial combination of column-vectors of the coordinates of $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_q$ with the same numbers $\lambda_1, \lambda_2, \dots, \lambda_q$ is equal to the zero column-vector and vice versa. ►

Theorem 6.5. *Let a basis in* **V** *comprise n vectors. Then any system of m (m > n) vectors in* **V** *is linearly dependent.*

◄ By Theorem 6.3 it suffices to consider the case of $m = n + 1$.

Let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n+1}$ be arbitrary vectors in **V**. Expanding these vectors relative to the basis $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n)$, we have

$$\begin{aligned} \mathbf{x}_1 &= \xi_1^1\mathbf{e}_1 + \xi_1^2\mathbf{e}_2 + \dots + \xi_1^n\mathbf{e}_n \\ \mathbf{x}_2 &= \xi_2^1\mathbf{e}_1 + \xi_2^2\mathbf{e}_2 + \dots + \xi_2^n\mathbf{e}_n, \\ &\dots\dots\dots\dots\dots\dots\dots\dots \\ \mathbf{x}_{n+1} &= \xi_{n+1}^1\mathbf{e}_1 + \xi_{n+1}^2\mathbf{e}_2 + \dots + \xi_{n+1}^n\mathbf{e}_n. \end{aligned}$$

Arranging the coordinates of $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n+1}$ in matrix form, we obtain the $n \times (n + 1)$ matrix

$$\mathbf{K} = \begin{pmatrix} \xi_1^1 & \xi_2^1 & \dots & \xi_{n+1}^1 \\ \xi_1^2 & \xi_2^2 & \dots & \xi_{n+1}^2 \\ \dots & \dots & \dots & \dots \\ \xi_1^n & \xi_2^n & \dots & \xi_{n+1}^n \end{pmatrix},$$

where the jth column comprises the coordinates of $\mathbf{x}_j$ ($j = 1, 2, \dots, n + 1$).

Since the rank of **K** does not exceed the number n of rows, $(n + 1)$ columns in **K** are linearly dependent column-vectors being at the same time the column-vectors of the coordinates of $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n+1}$. Then by Theorem 6.4 the linear dependence of these column-vectors implies that the vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n+1}$ are linearly dependent. ►

Corollary. All bases in **V** comprise equal numbers of vectors.

◄ Let **e** be a basis comprising n vectors and let $\mathbf{e}'$ be a basis comprising n' vectors. By Theorem 6.5 the linear independence of $\mathbf{e}_1', \mathbf{e}_2', \dots, \mathbf{e}_{n'}'$ implies that $n' \leqslant n$. On the other hand, by Theorem 6.5 the linear independence of $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ implies that $n \leqslant n'$ Thus we have $n = n'$. ►

The ***dimension*** of a linear space **V** is the number of vectors contained in the basis of **V**.

Examples. (1) The basis of the real coordinate space $\mathbb{R}^n$ is formed by the vectors

$$\mathbf{e}_1 = (1, 0, \dots, 0, 0), \ \mathbf{e}_2 = (0, 1, \dots, 0, 0), \ \dots, \ \mathbf{e}_n = (0, 0, \dots, 0, 1).$$

◄ Indeed, the vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ are linearly independent since the identity

$$\alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2 + \dots + \alpha_n\mathbf{e}_n = \mathbf{0}$$

implies that

$$\alpha_1(1, 0, \ldots, 0, 0) + \alpha_2(0, 1, \ldots, 0, 0) + \ldots + \alpha_n(0, 0, \ldots, 0, 1) = (\alpha_1, \alpha_2, \ldots, \alpha_n) = (0, 0, \ldots, 0).$$

Whence $\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0$.

Besides, any vector $\xi = (\xi^1, \xi^2, \ldots, \xi^n)$ in $\mathbb{R}^n$ may be expressed as a linear combination of $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$, i.e.,

$$\xi = \xi^1(1, 0, \ldots, 0, 0) + \xi^2(0, 1, \ldots, 0, 0) + \ldots + \xi^n(0, 0, \ldots, 0, 1) = (\xi^1, \xi^2, \ldots, \xi^n).$$

This means that the dimension of $\mathbb{R}^n$ is equal to n. ►

(2) Recall that the homogeneous linear system

$$\begin{aligned} \alpha_{11}x_1 + \alpha_{12}x_2 + \ldots + \alpha_{1n}x_n &= 0, \\ \alpha_{21}x_1 + \alpha_{22}x_2 + \ldots + \alpha_{2n}x_n &= 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \alpha_{m1}x_1 + \alpha_{m2}x_2 + \ldots + \alpha_{mn}x_n &= 0, \end{aligned}$$

with nontrivial (nonzero) solutions possesses the fundamental system of solutions which forms a linear space. The dimension of this linear space of solutions of a homogeneous linear system is equal to the number of the fundamental solutions, i.e., to $n - r$, where r is the rank of the coefficient matrix and n is the number of the unknowns.

(3) The dimension of the linear space $\mathbf{M}_n$ of polynomials of degree not exceeding n is equal to $n + 1$.

◄ Since any polynomial of a degree not exceeding n takes the form

$$P(t) = \alpha_0 + \alpha_1 t + \ldots + \alpha_n t^n,$$

it suffices to show that the vectors $\mathbf{e}_1 = 1$, $\mathbf{e}_2 = t$, ..., $\mathbf{e}_{n+1} = t^n$ are linearly independent.

Consider the identity

$$\alpha_0 1 + \alpha_1 t + \ldots + \alpha_n t^n = 0, \tag{6.6}$$

where t is an arbitrary quantity. Setting $t = 0$, we have $\alpha_0 = 0$.

On differentiating (6.6) with respect to t, we have

$$\alpha_1 + 2\alpha_2 t + \ldots + n\alpha_n t^{n-1} = 0.$$

Whence, setting $t = 0$, we obtain $\alpha_1 = 0$.

Repeating this process, we arrive at the identity

$$\alpha_0 = \alpha_1 = \ldots = \alpha_n = 0.$$

This means that the vectors $\mathbf{e}_1 = 1$, $\mathbf{e}_2 = t$, ..., $\mathbf{e}_{n+1} = t^n$ are linearly independent. Hence, the dimension of $\mathbf{M}_n$ is equal to $n + 1$. ►

A linear space of dimension n is called the *n-dimensional linear space.* In symbols, we write $\dim \mathbf{V} = n$.

By way of convention the dimension of the linear space $\mathbf{V}$ is considered equal to n everywhere through this chapter unless otherwise mentioned.

It is clear that for any linear subspace $\mathbf{W}$ in $\mathbf{V}$ $\dim \mathbf{W} \leqslant n$.

We show that the linear space $\mathbf{V}$ comprises linear subspaces of any dimension $k \leqslant n$.

◄ Let $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ be a basis in $\mathbf{V}$. It is easy to verify that the linear span

$$\mathbf{W}_k = \mathbf{L}(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k)$$

is a linear subspace of dimension k. ►

By definition $\dim \{\boldsymbol{\theta}\} = 0$.

Theorem 6.6. *Let* $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k$ *be linearly independent vectors in a linear space* $\mathbf{V}$ *and* $k < n$. *Then there exist vectors* $\mathbf{a}_{k+1}, \mathbf{a}_{k+2}, \ldots, \mathbf{a}_n$ *in* $\mathbf{V}$ *such that the vectors* $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k, \mathbf{a}_{k+1}, \ldots, \mathbf{a}_n$ *form a basis in* $\mathbf{V}$.

◄ Let $\mathbf{b}$ be an arbitrary vector in $\mathbf{V}$. If the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k, \mathbf{b}$ are linearly dependent then

$$\mathbf{b} = \alpha_1\mathbf{a}_1 + \alpha_2\mathbf{a}_2 + \ldots + \alpha_k\mathbf{a}_k \tag{6.7}$$

because the linear independence of $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k$ implies that the nontrivial linear combination

$$\lambda_1\mathbf{a}_1 + \lambda_2\mathbf{a}_2 + \ldots + \lambda_k\mathbf{a}_k + \mu\mathbf{b} = \boldsymbol{\theta}$$

involves $\mu \neq 0$.

Suppose that each vector $\mathbf{b}$ in $\mathbf{V}$ admits a representation of the form (6.7). In this case the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k$ form a basis in $\mathbf{V}$ by definition. However, this contradicts to the fact that the number k of the vectors is smaller than the dimension n of $\mathbf{V}$. This means that there must exist a vector $\mathbf{a}_{k+1}$ in $\mathbf{V}$ such that the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k, \mathbf{a}_{k+1}$ are linearly independent.

If $k + 1 = n$ the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k, \mathbf{a}_{k+1}$ form a basis in $\mathbf{V}$.

If $k + 1 < n$ we repeat the previous reasoning for the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k, \mathbf{a}_{k+1}$.

This process enables us to complement the collection of the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k$ by the vectors $\mathbf{a}_{k+1}, \mathbf{a}_{k+2}, \ldots, \mathbf{a}_n$ so that these n vectors form a basis in $\mathbf{V}$. In other words, we may always construct a basis of $\mathbf{V}$ which incorporates the basis of a given subspace of $\mathbf{V}$. ►

Example. Construct the basis of $\mathbb{R}^4$ by complementing the collection of the vectors $\mathbf{a}_1 = (1, 2, 0, 1)$ and $\mathbf{a}_2 = (-1, 1, 1, 0)$.

◄ We choose the vectors $\mathbf{a}_3 = (1, 0, 0, 0)$ and $\mathbf{a}_4 = (0, 1, 0, 0)$ in $\mathbb{R}^4$. Let us show that the vectors $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4$ form a basis of $\mathbb{R}^4$.

Consider the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 & 1 \\ -1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

whose rows comprise the coordinates of $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4$.

Notice that the rank of $\mathbf{A}$ is equal to 4. This means that the rows of $\mathbf{A}$ and, consequently, the vectors $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4$ are linearly independent. Thus they form a basis of $\mathbb{R}^4$. ►

We proceed similarly in the general case.

Suppose we wish to construct the basis of $\mathbb{R}^n$ by complementing the collection of k linearly independent vectors

$$\mathbf{a}_1 = (\alpha_{11}, \alpha_{12}, ..., \alpha_{1n}), \quad \mathbf{a}_2 = (\alpha_{21}, \alpha_{22}, ..., \alpha_{2n}),$$
$$..., \mathbf{a}_k = (\alpha_{k1}, \alpha_{k2}, ..., \alpha_{kn}).$$

By elementary row operations the matrix

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{k1} & \alpha_{k2} & \cdots & \alpha_{kn} \end{pmatrix}$$

can be reduced to the schematic form and then complemented by $n - k$ rows of the form

$$(0, 0, ..., 1, 0, ..., 0)$$

so that the rank of the matrix obtained becomes equal to n.

Theorem 6.7. *Let* $\mathbf{W}_1$ *and* $\mathbf{W}_2$ *be linear subspaces in* $\mathbf{V}$. *Then*

$$\dim(\mathbf{W}_1 + \mathbf{W}_2) + \dim(\mathbf{W}_1 \cap \mathbf{W}_2) = \dim \mathbf{W}_1 + \dim \mathbf{W}_2.$$

6.5 Changing a Basis

Let $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n)$ and $\mathbf{e}' = (\mathbf{e}_1', \mathbf{e}_2', ..., \mathbf{e}_n')$ be bases of $\mathbf{V}$. Expanding the vectors in $\mathbf{e}'$ with respect to the basis $\mathbf{e}$, we have

$$\mathbf{e}_j' = \sum_{i=1}^{n} \alpha_j^i \mathbf{e}_i, \quad j = 1, 2, ..., n \tag{6.8}$$

or

$$\mathbf{e}_j' = \alpha_j^1 \mathbf{e}_1 + \alpha_j^2 \mathbf{e}_2 + ... + \alpha_j^n \mathbf{e}_n, \quad j = 1, 2, ..., n.$$

In matrix form, (6.8) becomes

$$(\mathbf{e}'_1, \mathbf{e}'_2, \ldots, \mathbf{e}'_n) = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)\begin{pmatrix} \alpha_1^1 & \alpha_2^1 & \ldots & \alpha_n^1 \\ \alpha_1^2 & \alpha_2^2 & \ldots & \alpha_n^2 \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_1^n & \alpha_2^n & \ldots & \alpha_n^n \end{pmatrix}. \qquad (6.9)$$

The matrix

$$\mathbf{A} = \begin{pmatrix} \alpha_1^1 & \alpha_2^1 & \ldots & \alpha_n^1 \\ \alpha_1^2 & \alpha_2^2 & \ldots & \alpha_n^2 \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_1^n & \alpha_2^n & \ldots & \alpha_n^n \end{pmatrix}$$

describes the transition from the basis **e** to the basis **e**′ and is called the *transition matrix.*

Properties of transition matrices. (a) The determinant of **A** is not equal to zero, i.e., det $\mathbf{A} \neq 0$.

Let us assume the converse, i.e., det $\mathbf{A} = 0$. This means that the columns in **A** are linearly dependent.

◀ Since the column-vectors in **A** are the column-vectors of the coordinates of $\mathbf{e}'_1, \mathbf{e}'_2, \ldots, \mathbf{e}'_n$ relative to the basis **e**, Theorem 6.4 implies that $\mathbf{e}'_1, \mathbf{e}'_2, \ldots, \mathbf{e}'_n$ are linearly dependent vectors. However, this contradicts to the fact that **e**′ is the basis of **V**. Consequently, the assumption that det $\mathbf{A} = 0$ is false. Hence, det $\mathbf{A} \neq 0$. ▶

(b) If $\xi^1, \xi^2, \ldots, \xi^n$ and $\xi'^1, \xi'^2, \ldots, \xi'^n$ are the coordinates of a vector **x** relative to the bases **e** and **e**′, respectively, then

$$\begin{pmatrix} \xi^1 \\ \xi^2 \\ \vdots \\ \xi^n \end{pmatrix} = \mathbf{A}\begin{pmatrix} \xi'^1 \\ \xi'^2 \\ \vdots \\ \xi'^n \end{pmatrix}. \qquad (6.10)$$

◀ Substituting (6.8) into the formula

$$\mathbf{x} = \sum_{i=1}^{n} \xi^i \mathbf{e}_i = \sum_{j=1}^{n} \xi'^j \mathbf{e}'_j$$

we obtain

$$\sum_{i=1}^{n} \xi^i \mathbf{e}_i = \sum_{j=1}^{n} \xi'^j \left(\sum_{i=1}^{n} \alpha_j^i \mathbf{e}_i \right) = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \alpha_j^i \xi'^j \right) \mathbf{e}_i.$$

Since, relative to a given basis, the expansion of a vector is unique, the above relation gives

$$\xi^i = \sum_{j=1}^{n} \alpha_j^i \xi'^j, \quad i = 1, 2, \ldots, n.$$

Arranging these expressions in matrix form, we easily obtain (6.10). ►

(c) The inverse $\mathbf{A}^{-1}$ of $\mathbf{A}$ is the transition matrix from the basis $\mathbf{e}'$ to the basis $\mathbf{e}$.

◄ To verify this property it suffices to post-multiply (6.9) by $\mathbf{A}^{-1}$. ►

6.6 Euclidean Spaces

A real linear space $\mathbf{V}$ is called a *real Euclidean space* if, given two arbitrary vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbf{V}$ there exists a number $(\mathbf{x}, \mathbf{y})$ (or $\mathbf{x}\cdot\mathbf{y}$) such that for any vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ in $\mathbf{V}$ and any arbitrary number α the following conditions are satisfied:

(a) $(\mathbf{x}, \mathbf{y}) = (\mathbf{y}, \mathbf{x})$;

(b) $(\mathbf{x} + \mathbf{y}, \mathbf{z}) = (\mathbf{x}, \mathbf{z}) + (\mathbf{y}, \mathbf{z})$;

(c) $(\alpha\mathbf{x}, \mathbf{y}) = \alpha(\mathbf{x}, \mathbf{y})$;

(d) $(\mathbf{x}, \mathbf{x}) \geqslant 0$; the identity holds if and only if $\mathbf{x} = \boldsymbol{\theta}$.

The number $(\mathbf{x}, \mathbf{y})$ is called the *scalar product* of vectors $\mathbf{x}$ and $\mathbf{y}$.

Examples of Euclidean spaces. (1) The space $\mathbf{V}_3$ of free vectors with the scalar product of $\mathbf{a}$ and $\mathbf{b}$ defined as

$$(\mathbf{a}, \mathbf{b}) = |\mathbf{a}||\mathbf{b}| \cos (\widehat{\mathbf{a}, \mathbf{b}}).$$

(2) The real coordinate space $\mathbb{R}^n$ with the scalar product of two vectors $\xi = (\xi^1, \xi^2, ..., \xi^n)$ and $\eta = (\eta^1, \eta^2, ..., \eta^n)$ in $\mathbb{R}^n$ given by

$$(\xi, \eta) = \sum_{i=1}^{n} \xi^i \eta^i.$$

(3) A linear subspace of the Euclidean space.

In all these cases the validity of conditions (a)-(d) are directly verified.

Employing the definition of the Euclidean space we can easily prove the following properties:

(1) $(\boldsymbol{\theta}, \mathbf{x}) = 0$;

(2) $(\mathbf{x}, \alpha\mathbf{y}) = \alpha(\mathbf{x}, \mathbf{y})$;

(3) $(\mathbf{x}, \mathbf{y} + \mathbf{z}) = (\mathbf{x}, \mathbf{y}) + (\mathbf{x}, \mathbf{z})$;

(4) $\left(\sum_{i=1}^{l} \alpha_i \mathbf{x}_i, \sum_{j=1}^{m} \beta_j \mathbf{y}_j\right) = \sum_{i=1}^{l} \sum_{j=1}^{m} \alpha_i \beta_j (\mathbf{x}_i, \mathbf{y}_j)$.

Theorem 6.8. *For any two vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *of the Euclidean space there holds the inequality*

$$(\mathbf{x}, \mathbf{y})^2 \leqslant (\mathbf{x}, \mathbf{x})(\mathbf{y}, \mathbf{y})$$

(*the Cauchy-Schwarz inequality*).

◄ If $(\mathbf{x}, \mathbf{x}) = 0$, then $\mathbf{x} = \boldsymbol{\theta}$ and the inequality holds since $(\boldsymbol{\theta}, \mathbf{y}) = 0$.

Let $(\mathbf{x}, \mathbf{x}) \neq 0$. Then $(\mathbf{x}, \mathbf{x}) > 0$.

By the definition of the scalar product the inequality

$$(t\mathbf{x} - \mathbf{y},\ t\mathbf{x} - \mathbf{y}) \geqslant 0 \tag{6.11}$$

holds for any $\mathbf{x}$ and $\mathbf{y}$ in $\mathbf{V}$ and for any real number t.

The inequality (6.11) may be written as

$$t^2(\mathbf{x}, \mathbf{x}) - 2t(\mathbf{x}, \mathbf{y}) + (\mathbf{y}, \mathbf{y}) \geqslant 0.$$

The left side of this inequality can be considered as a quadratic trinomial with respect to t. Since the sign of the trinomial remains unchanged for any t we infer that the discriminant of the trinomial is nonpositive, i.e.,

$$(\mathbf{x}, \mathbf{y})^2 - (\mathbf{x}, \mathbf{x})(\mathbf{y}, \mathbf{y}) \leqslant 0.$$

Whence we have the desired inequality. ►

Definition. A *length* of an arbitrary vector $\mathbf{x}$ is a number $|\mathbf{x}|$ given by $x = \sqrt{(\mathbf{x}, \mathbf{x})}$.

It is easily seen that $|\mathbf{x}| \geqslant 0$ for any $\mathbf{x}$, and $|\mathbf{x}| = 0$ only if $\mathbf{x} = \boldsymbol{\theta}$.

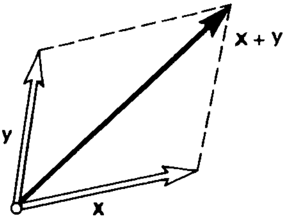


Fig. 6.7

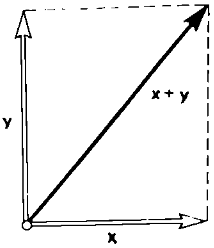


Fig. 6.8.

Consider the following identities

$$\begin{aligned} |\mathbf{x} + \mathbf{y}|^2 &= (\mathbf{x} + \mathbf{y},\ \mathbf{x} + \mathbf{y}) \\ &= (\mathbf{x}, \mathbf{x}) + 2(\mathbf{x}, \mathbf{y}) + (\mathbf{y}, \mathbf{y}) = |\mathbf{x}|^2 + 2(\mathbf{x}, \mathbf{y}) + |\mathbf{y}|^2. \end{aligned} \tag{6.12}$$

Replacing $2(\mathbf{x}, \mathbf{y})$ by $2|(\mathbf{x}, \mathbf{y})|$ and applying the Cauchy-Schwarz inequality $|(\mathbf{x}, \mathbf{y})| \leqslant |\mathbf{x}||\mathbf{y}|$, we have

$$|\mathbf{x} + \mathbf{y}|^2 \leqslant |\mathbf{x}|^2 + 2|\mathbf{x}||\mathbf{y}| + |\mathbf{y}|^2 = (|\mathbf{x}| + |\mathbf{y}|)^2.$$

Extracting the square root, we arrive at the *triangle inequality* (Fig. 6.7)

$$|\mathbf{x} + \mathbf{y}| \leqslant |\mathbf{x}| + |\mathbf{y}|.$$

The *angle* between the nonzero vectors $\mathbf{x}$ and $\mathbf{y}$ is the number φ such that $0 \leqslant \varphi \leqslant \pi$ and

$$\cos\varphi = \frac{(\mathbf{x}, \mathbf{y})}{|\mathbf{x}|\,|\mathbf{y}|}.$$

By Theorem 6.8 it follows that

$$-1 \leqslant \frac{(\mathbf{x}, \mathbf{y})}{|\mathbf{x}|\,|\mathbf{y}|} \leqslant 1$$

for any nonzero vectors $\mathbf{x}$ and $\mathbf{y}$, i.e., the range of $\cos \varphi$ completely corresponds to its domain. In other words, the angle between any two vectors is correctly defined.

The vectors $\mathbf{x}$ and $\mathbf{y}$ are called *orthogonal* if $(\mathbf{x}, \mathbf{y}) = 0$.

From (6.12) it follows that, given orthogonal vectors, there holds the identity

$$|\mathbf{x} + \mathbf{y}|^2 = |\mathbf{x}|^2 + |\mathbf{y}|^2$$

which may be thought as generalization of the theorem of Pythagoras, namely, the square of the lengths of two orthogonal vectors is equal to the sum of the squares of the lengths of these vectors (Fig. 6.8).

The collection of vectors $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_k$ is called *orthogonal* if $(\mathbf{f}_i, \mathbf{f}_j) = 0$ for $i \neq j$, and *orthonormal* if

$$(\mathbf{f}_i, \mathbf{f}_j) = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j. \end{cases}$$

Theorem 6.9. *Any orthonormal collection of vectors* $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_k$ *is linearly independent.*

◄ Computing the scalar products of

$$\alpha_1\mathbf{f}_1 + \alpha_2\mathbf{f}_2 + \ldots + \alpha_k\mathbf{f}_k = \boldsymbol{\theta}$$

by $\mathbf{f}_j$ $(j = 1, 2, \ldots, k)$, we have k identities

$$\alpha_j(\mathbf{f}_j, \mathbf{f}_j) = 0 \quad (j = 1, 2, \ldots, k).$$

Since $(\mathbf{f}_j, \mathbf{f}_j) = 1$ then $\alpha_j = 0$ for every j $(j = 1, 2, \ldots, k)$. ►

6.7 Orthogonalization

Let $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_k$ be linearly independent vectors in a real Euclidean space. We describe the procedure of constructing a collection of k orthogonal vectors using $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_k$ (Fig. 6.9).

◄ Set $\mathbf{g}_1 = \mathbf{f}_1$. For the vector $\mathbf{g}_2 = \mathbf{f}_2 - \alpha_1\mathbf{g}_1$ to be orthogonal to $\mathbf{g}_1$ it is necessary that the identity

$$0 = (\mathbf{f}_2, \mathbf{g}_1) - \alpha_1(\mathbf{g}_1, \mathbf{g}_1)$$

holds.

Whence

$$\alpha_1 = \frac{(\mathbf{f}_2, \mathbf{g}_1)}{(\mathbf{g}_1, \mathbf{g}_1)}\,.$$

Thus the vector

$$\mathbf{g}_2 = \mathbf{f}_2 - \frac{(\mathbf{f}_2, \mathbf{g}_1)}{(\mathbf{g}_1, \mathbf{g}_1)} \mathbf{g}_1$$

is orthogonal to $\mathbf{g}_1$.

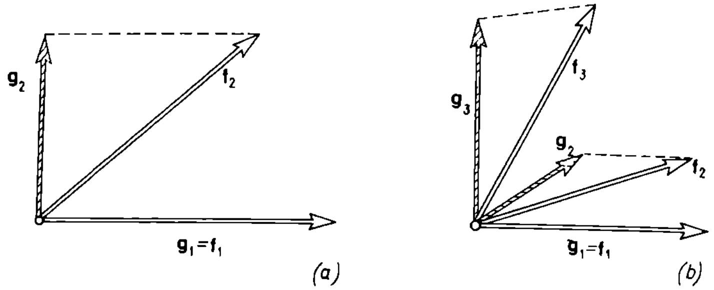


Fig. 6.9

Using $\mathbf{g}_1$, $\mathbf{g}_2$, $\mathbf{f}_3$, we construct the vector $\mathbf{g}_3 = \mathbf{f}_3 - \beta_1\mathbf{g}_1 - \beta_2\mathbf{g}_2$ which is orthogonal both to $\mathbf{g}_1$ and to $\mathbf{g}_2$. It suffices to require that the numbers β_1 and β_2 satisfy the following conditions

$$0 = (\mathbf{f}_3, \mathbf{g}_1) - \beta_1(\mathbf{g}_1, \mathbf{g}_1) \quad \text{and} \quad 0 = (\mathbf{f}_3, \mathbf{g}_2) - \beta_2(\mathbf{g}_2, \mathbf{g}_2).$$

Whence

$$\beta_1 = \frac{(\mathbf{f}_3, \mathbf{g}_1)}{(\mathbf{g}_1, \mathbf{g}_1)} \quad \text{and} \quad \beta_2 = \frac{(\mathbf{f}_3, \mathbf{g}_2)}{(\mathbf{g}_2, \mathbf{g}_2)},$$

so that the vector

$$\mathbf{g}_3 = \mathbf{f}_3 - \frac{(\mathbf{f}_3, \mathbf{g}_1)}{(\mathbf{g}_1, \mathbf{g}_1)} \mathbf{g}_1 - \frac{(\mathbf{f}_3, \mathbf{g}_2)}{(\mathbf{g}_2, \mathbf{g}_2)} \mathbf{g}_2$$

is orthogonal both to $\mathbf{g}_1$ and to $\mathbf{g}_2$.

Similarly, the vector

$$\mathbf{g}_i = \mathbf{f}_i - \frac{(\mathbf{f}_i, \mathbf{g}_1)}{(\mathbf{g}_1, \mathbf{g}_1)} \mathbf{g}_1 - \frac{(\mathbf{f}_i, \mathbf{g}_2)}{(\mathbf{g}_2, \mathbf{g}_2)} \mathbf{g}_2 - \dots - \frac{(\mathbf{f}_i, \mathbf{g}_{i-1})}{(\mathbf{g}_{i-1}, \mathbf{g}_{i-1})} \mathbf{g}_{i-1}$$
$$(i = 3, 4, \dots, k)$$

is orthogonal to each of the vectors $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_{i-1}$.

Therefore the vectors $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_k$ form a collection of orthogonal vectors.

Dividing each vector $\mathbf{g}_i$ $(i = 1, 2, \dots, k)$ by its length $|\mathbf{g}_i|$ we obtain the collection of the orthonormal vectors (see Fig. 6.10)

$$\mathbf{e}_1 = \frac{\mathbf{g}_1}{|\mathbf{g}_1|}, \quad \mathbf{e}_2 = \frac{\mathbf{g}_2}{|\mathbf{g}_2|}, \quad \ldots, \quad \mathbf{e}_k = \frac{\mathbf{g}_k}{|\mathbf{g}_k|}. \blacktriangleright$$

The basis $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ of the Euclidean space is called *orthonormal* if

$$(\mathbf{e}_i, \mathbf{e}_j) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad (i, j = 1, 2, \ldots, n).$$

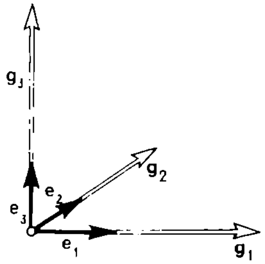


Fig. 6.10

Summing up, we arrive at the following theorem.

Theorem 6.10. *There exists an orthonormal basis in any Euclidean space.*

Example. Let $\mathbf{a}_1 = (1, 2, 0, 1)$, $\mathbf{a}_2 = (-1, 1, 1, 0)$ and $\mathbf{a}_3 = (2, 2, 0, 1)$ form a basis in a Euclidean space.

◀ We apply the procedure of orthogonalization to construct the orthonormal basis in this space.

Set $\mathbf{b}_1 = \mathbf{a}_1$ and $\mathbf{b}_2 = \mathbf{a}_2 - \alpha\mathbf{a}_1$. For the vector $\mathbf{b}_2$ to be orthogonal to the vector $\mathbf{a}_1$ it is necessary that the identity

$$0 = (\mathbf{a}_2, \mathbf{a}_1) - \alpha(\mathbf{a}_1, \mathbf{a}_1)$$

is fulfilled.

Whence

$$\alpha = \frac{(\mathbf{a}_2, \mathbf{a}_1)}{(\mathbf{a}_1, \mathbf{a}_1)} = \frac{-1 + 2}{1 + 4 + 1} = \frac{1}{6},$$

so that

$$\mathbf{b}_2 = (-1, 1, 1, 0) - \frac{1}{6}(1, 2, 0, 1) = \left(-\frac{7}{6}, \frac{2}{3}, 1, -\frac{1}{6}\right).$$

Evidently, the vector $\mathbf{c}_2 = 6\mathbf{b}_2 = (-7, 4, 6, -1)$ is orthogonal to the vector $\mathbf{a}_1$. Indeed,

$$(\mathbf{c}_2, \mathbf{a}_1) = (-7) \times 1 + 4 \times 2 + 6 \times 0 + (-1) \times 1 = 0.$$

For vector $\mathbf{b}_3 = \mathbf{a}_3 - \beta\mathbf{a}_1 - \gamma\mathbf{c}_2$ to be orthogonal both to $\mathbf{a}_1$ and to $\mathbf{c}_2$ it is necessary that the identities

$$0 = (\mathbf{a}_3, \mathbf{a}_1) - \beta(\mathbf{a}_1, \mathbf{a}_1) \text{ and } 0 = (\mathbf{a}_3, \mathbf{c}_2) - \gamma(\mathbf{c}_2, \mathbf{c}_2)$$

are fulfilled.

Whence

$$\beta = \frac{(\mathbf{a}_3, \mathbf{a}_1)}{(\mathbf{a}_1, \mathbf{a}_1)} = \frac{2 + 4 + 1}{6} = \frac{7}{6}$$

and

$$\gamma = \frac{(\mathbf{a}_3, \mathbf{c}_2)}{(\mathbf{c}_2, \mathbf{c}_2)} = \frac{-14 + 8 - 1}{49 + 16 + 36 + 1} = -\frac{7}{102}$$

so that

$$\begin{aligned}\mathbf{b}_3 &= (2, 2, 0, 1) - \frac{7}{6}(1, 2, 0, 1) + \frac{7}{102}(-7, 4, 6, -1)\\ &= \left(\frac{204 - 119 - 49}{102}, \frac{204 - 238 + 28}{102}, \frac{42}{102}, \frac{102 - 119 - 7}{102}\right)\\ &= \left(\frac{6}{17}, -\frac{1}{17}, \frac{7}{17}, -\frac{4}{17}\right).\end{aligned}$$

It is easy to see that the vector $\mathbf{c}_3 = 17\mathbf{b}_3 = (6, -1, 7, -4)$ is orthogonal both to $\mathbf{a}_1$ and to $\mathbf{c}_2$.

The vectors $\mathbf{a}_1, \mathbf{c}_2, \mathbf{c}_3$ form a collection of orthogonal vectors. Dividing $\mathbf{a}_1, \mathbf{c}_2, \mathbf{c}_3$ by their lengths $\sqrt{6}$, $\sqrt{102}$, $\sqrt{102}$, respectively, we arrive at the collection of the orthonormal vectors

$$\mathbf{e}_1 = \frac{1}{\sqrt{6}}(1, 2, 0, 1),$$

$$\mathbf{e}_2 = \frac{1}{\sqrt{102}}(-7, 4, 6, 1),$$

$$\mathbf{e}_3 = \frac{1}{\sqrt{102}}(6, -1, 7, 4).$$

The vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ form an orthonormal basis of the Euclidean space. ►

If we wish to compute the scalar product of two vectors in a Euclidean space it is helpful to expand the vectors relative to the orthonormal basis because in this case the desired product can be expressed by the simplest formula.

◄ Let $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ be an orthonormal basis of a Euclidean space. Computing the scalar product of the vectors

$$\mathbf{x} = \sum_{i=1}^{n} \xi^i \mathbf{e}_i \quad \text{and} \quad \mathbf{y} = \sum_{j=1}^{n} \eta^j \mathbf{e}_j$$

we have

$$(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} \xi^i \mathbf{e}_i, \quad \sum_{j=1}^{n} \eta^j \mathbf{e}_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \xi^i \eta^j (\mathbf{e}_i, \mathbf{e}_j) = \sum_{i=1}^{n} \xi^i \eta^i.$$

In particular

$$(\mathbf{x}, \mathbf{x}) = \sum_{i=1}^{n} (\xi^i)^2.$$

Whence

$$|\mathbf{x}| = \sqrt{\sum_{i=1}^{n} (\xi^i)^2}. \quad ►$$

6.8 Orthocompliments of Linear Subspaces

Suppose that we are given a linear subspace **W** in a Euclidean space **V**.

A collection of vectors in **V** that satisfy the condition

$$(\mathbf{y}, \mathbf{x}) = 0,$$

where $\mathbf{x}$ is an arbitrary vector in **W**, is said to be an *orthocompliment* of **W**, denoted by $\mathbf{W}^{\perp}$. In other words, the orthocompliment $\mathbf{W}^{\perp}$ of **W** consists of vectors $\mathbf{y}$ in **V** that are orthogonal to all vectors in **W**.

Properties of orthocompliments. (a) The orthocompliment $\mathbf{W}^{\perp}$ of the linear subspace **W** is a linear subspace in **V**.

◄ Let $\mathbf{y}_1$ and $\mathbf{y}_2$ belong to $\mathbf{W}^{\perp}$, i.e.,

$$(\mathbf{y}_1, \mathbf{x}) = 0 \quad \text{and} \quad (\mathbf{y}_2, \mathbf{x}) = 0$$

for any vector $\mathbf{x}$ in **W**.

Adding these expressions, by properties of scalar products, we derive that for any vector $\mathbf{x}$ in **W** there holds

$$(\mathbf{y}_1 + \mathbf{y}_2, \mathbf{x}) = 0.$$

Whence we conclude that $\mathbf{y}_1 + \mathbf{y}_2$ belongs to $\mathbf{W}^{\perp}$.

Since $(\mathbf{y}, \mathbf{x}) = 0$ for any vector $\mathbf{x}$ in **W**, we infer that $(\alpha\mathbf{y}, \mathbf{x}) = \alpha(\mathbf{y}, \mathbf{x})$ and, consequently, $\alpha\mathbf{y}$ is a vector in $\mathbf{W}^{\perp}$. ►

(b) $\mathbf{V} = \mathbf{W} \oplus \mathbf{W}^{\perp}$.

◄ This follows from the fact that the zero vector is the only vector contained both in **W** and in $\mathbf{W}^\perp$. ►

Property (b) implies that any vector **x** in **V** is uniquely expressed as

$$\mathbf{x} = \mathbf{y} + \mathbf{z}, \qquad (*)$$

where **y** is a vector in **W** and **z** is a vector in $\mathbf{W}^\perp$ (Fig. 6.11).

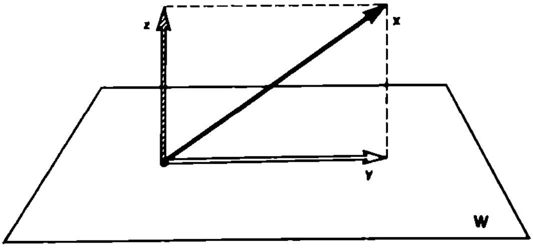


Fig. 6.11

The vector **y** in **W** is called the **W**-*component* of **x** and **z** in $\mathbf{W}^\perp$ the *orthogonal component* of **x** with respect to **W**.

Let **x** be a given vector in a linear subspace **W**. We show how to construct the **W**-component of **x** and the orthogonal component of **x** with respect to **W**.

◄ We may assume that the orthonormal basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k$ is introduced in **W**. Then the **W**-component of **x**, denoted by **y**, is expressed as

$$\mathbf{y} = \alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2 + \ldots + \alpha_k\mathbf{e}_k.$$

Substituting **y** into (*), we have

$$\mathbf{x} = \alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2 + \ldots + \alpha_k\mathbf{e}_k + \mathbf{z},$$

where **z** is the desired orthogonal component of **x** with respect to **W**.

Recall that **z** is orthogonal to **W**. Then, multiplying **x** successively by $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k$, we obtain

$$(\mathbf{x}, \mathbf{e}_1) = \alpha_1, \ (\mathbf{x}, \mathbf{e}_2) = \alpha_2, \ \ldots, \ (\mathbf{x}, \mathbf{e}_k) = \alpha_k.$$

The vectors

$$\mathbf{y} = (\mathbf{x}, \mathbf{e}_1)\mathbf{e}_1 + (\mathbf{x}, \mathbf{e}_2)\mathbf{e}_2 + \ldots + (\mathbf{x}, \mathbf{e}_k)\mathbf{e}_k$$

and

$$\mathbf{z} = \mathbf{x} - \mathbf{y}$$

are the **W**-component of **x** and the orthogonal component of **x** with respect to **W**, correspondingly. ►

Example. Construct the **W**-component of **x** = (4, 2, 3, 5) for the linear subspace $\mathbf{W} \subset \mathbb{R}^4$, specified by the system of equations

$$\begin{cases} x_1 + x_2 + x_3 + x_4 = 0, \\ x_1 - x_2 - x_3 + x_4 = 0. \end{cases}$$

◄ The vectors $\mathbf{a}_1 = (1, 0, 0, -1)$ and $\mathbf{a}_2 = (0, 1, -1, 0)$ form the fundamental system of solutions and, consequently, the basis of **W**. To construct the orthonormal basis of **W** it suffices to divide $\mathbf{a}_1$ and $\mathbf{a}_2$ by their lengths. We obtain

$$\mathbf{e}_1 = \left(\frac{1}{\sqrt{2}}, 0, 0, -\frac{1}{\sqrt{2}}\right) \quad \text{and} \quad \mathbf{e}_2 = \left(0, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0\right).$$

The vector

$$\mathbf{y} = \left(\frac{4}{\sqrt{2}} - \frac{5}{\sqrt{2}}\right)\left(\frac{1}{2}, 0, 0, -\frac{1}{\sqrt{2}}\right)$$
$$+ \left(\frac{2}{\sqrt{2}} - \frac{3}{\sqrt{2}}\right)\left(0, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0\right)$$
$$= \left(-\frac{1}{2}, 0, 0, \frac{1}{2}\right) + \left(0, -\frac{1}{2}, \frac{1}{2}, 0\right) = \frac{1}{2}(-1, -1, 1, 1)$$

is the **W**-component of **x** = (4, 2, 3, 5) and the vector

$$\mathbf{z} = \mathbf{x} - \mathbf{y} = \left(4\frac{1}{2}, 2\frac{1}{2}, 2\frac{1}{2}, 4\frac{1}{2}\right)$$

is the orthogonal component of **x** with respect to **W**. ►

6.9 Unitary Spaces

A *unitary space* is a complex linear space **U** such that given any ordered pair of vectors **x** and **y** in **U**, there exists a number called the *scalar product* (**x**, **y**) of **x** and **y** so that for any vectors **x**, **y** and **z** in **U** the following conditions are satisfied

(a) $(\mathbf{y}, \mathbf{x}) = \overline{(\mathbf{x}, \mathbf{y})}$ (here $\overline{(\mathbf{x}, \mathbf{y})}$ is the complex conjugate of (**x**, **y**));

(b) $(\mathbf{x} + \mathbf{y}, \mathbf{z}) = (\mathbf{x}, \mathbf{z}) + (\mathbf{y}, \mathbf{z})$;

(c) $(\alpha\mathbf{x}, \mathbf{y}) = \alpha(\mathbf{x}, \mathbf{y})$;

(d) $(\mathbf{x}, \mathbf{y}) \geqslant 0$, the identity $(\mathbf{x}, \mathbf{x}) = 0$ holding for $\mathbf{x} = \boldsymbol{\theta}$ only.

Example. Consider a coordinate space $\mathbb{C}^n$ whose elements are all possible ordered *n*-tuples of *n* complex numbers. The scalar product of the vectors $\boldsymbol{\xi} = (\xi^1, \xi^2, \ldots, \xi^n)$ and $\boldsymbol{\eta} = (\eta^1, \eta^2, \ldots, \eta^n)$ in $\mathbb{C}^n$ is expressed as

$$(\boldsymbol{\xi}, \boldsymbol{\eta}) = \sum_{j=1}^{n} \xi^j \overline{\eta^j}.$$

6.10 Linear Mappings

Let **V** and **W** be two linear spaces, either both real or both complex.

A *linear mapping* of a linear space **V** into a linear space **W** is a rule or operation $\mathscr{A}$ which associates every vector $\mathbf{x}$ in **V** with a unique vector $\mathbf{y} = \mathscr{A}\mathbf{x}$ in **W** so that

(i) $\mathscr{A}(\mathbf{x}_1 + \mathbf{x}_2) = \mathscr{A}\mathbf{x}_1 + \mathscr{A}\mathbf{x}_2$ and (ii) $\mathscr{A}(\alpha\mathbf{x}) = \alpha\mathscr{A}\mathbf{x}$.

The conditions (i) and (ii) can be expressed in a single formula

$$\mathscr{A}(\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2) = \alpha_1\mathscr{A}\mathbf{x}_1 + \alpha_2\mathscr{A}\mathbf{x}_2.$$

In symbols, we write $\mathscr{A}: \mathbf{V} \to \mathbf{W}$ to designate a mapping of **V** into **W**.

To illustrate the concept of linear mapping we shall turn to some examples:

(1) Let $\mathbf{V} = \mathbf{W} = \mathbf{M}_n$, where $\mathbf{M}_n$ is a space of polynomials of degree not exceeding n. Laws of differentiation set every polynomial in $\mathbf{M}_n$ into correspondence with its derivative which is also a polynomial of degree not exceeding n, i.e., belongs to $\mathbf{M}_n$. Yet, the derivative of the sum of the polynomials (functions) is equal to the sum of the derivatives of these polynomials and the derivative of the polynomial multiplied by a scalar is equal to the derivative of the polynomial multiplied by the same scalar. Under the operation $\mathscr{D}: \mathbf{M}_n \to \mathbf{M}_n$ which associates every polynomial in $\mathbf{M}_n$ with its derivative the conditions (i) and (ii) are fulfilled. Hence, the operation $\mathscr{D}: \mathbf{M}_n \to \mathbf{M}_n$ is a linear mapping.

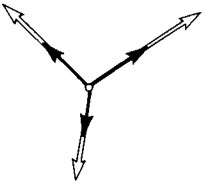


Fig. 6.12

(2) The rule that sets every vector $\mathbf{x}$ in **V** into correspondence with a vector $\lambda\mathbf{x}$ in **V**, where λ is a given scalar, is a linear mapping, called the *similarity* mapping (Fig. 6.12).

(3) Let $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ be a basis of **V**. The rule $\mathscr{P}: \mathbf{V} \to \mathbf{V}$, which associates every vector

$$\mathbf{x} = \xi^1\mathbf{e}_1 + \xi^2\mathbf{e}_2 + \ldots + \xi^k\mathbf{e}_k + \ldots + \xi^n\mathbf{e}_n = \sum_{i=1}^{n} \xi^i\mathbf{e}_i$$

in **V** with the vector

$$\mathscr{P}\mathbf{x} = \xi^1\mathbf{e}_1 + \xi^2\mathbf{e}_2 + \ldots + \xi^k\mathbf{e}_k = \sum_{i=1}^{k} \xi^i\mathbf{e}_i$$

in **V**, given $k < n$, is a linear mapping called the *projective* mapping. Hence, the linear mapping $\mathscr{P}$: **V** → **V** sets every vector **x** from **V** into correspondence with the vector $\mathscr{P}\mathbf{x}$ in **V** (Fig. 6.13).

(4) The set $\mathbf{T}_2$ of trigonometric polynomials of the form

$$\alpha \cos x + \beta \sin x$$

forms a linear space. Similarly to (1), it is easy to see that the rule

$$\mathscr{D} = \frac{d}{dx} : \alpha \cos x + \beta \sin x \to -\alpha \sin x + \beta \cos x$$

is a linear mapping

$$\mathscr{D}: \mathbf{T}_2 \to \mathbf{T}_2.$$

(5) Let **A** be a given $m \times n$ matrix from $\mathbb{R}^{m\times n}$ and let **X** be an $n \times 1$ matrix, i.e., a column-vector. Pre-multiplying **X** by **A** constitutes the linear mapping

$$\mathbf{X} \in \mathbb{R}^{n\times 1} \to \mathbf{AX} \in \mathbb{R}^{m\times 1}$$

which associates every column-vector **X** in $\mathbb{R}^{n\times 1}$ with a column-vector in $\mathbb{R}^{m\times 1}$.

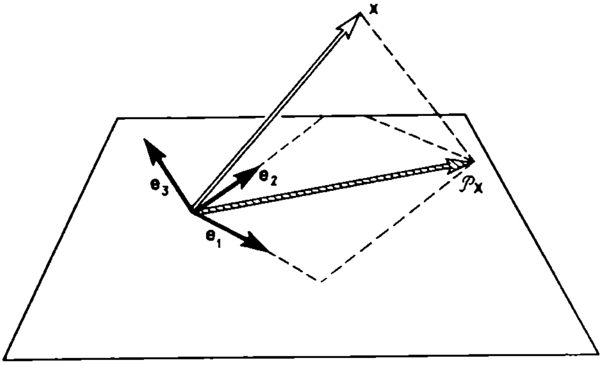


Fig. 6.13

The mapping $\mathscr{I}$: **V** → **V** such that $\mathscr{I}\mathbf{x} = \mathbf{x}$ is called the *identity mapping*.

An *image* of a linear mapping $\mathscr{A}$: **V** → **W** is a set im $\mathscr{A}$ of all vectors from **W** satisfying the following condition: a vector **y** in **W** is contained in im $\mathscr{A}$ if there exists a vector **x** in **V** such that $\mathscr{A}\mathbf{x} = \mathbf{y}$.

Referring to the examples of linear mappings we mention the folowing:

(1′) The image of the differential mapping $\mathscr{D}$: $\mathbf{M}_n \to \mathbf{M}_n$ is the set of polynomials of degree not exceeding $n - 1$.

(2′) The image of the similarity mapping coincides with the linear space $\mathbf{V}$.

(3′) The image of the projective mapping $\mathscr{P}$: $\mathbf{V} \to \mathbf{V}$ is the subspace

$$\mathbf{W}_k = \mathbf{L}(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k).$$

(4′) The image of the differential mapping $\mathscr{D}$: $\mathbf{T}_2 \to \mathbf{T}_2$ coincides with the linear space $\mathbf{T}_2$.

Proposition. The image of the linear mapping $\mathscr{A}$: $\mathbf{V} \to \mathbf{W}$ is a linear subspace of $\mathbf{W}$.

◀ Let $\mathbf{y}_1$ and $\mathbf{y}_2$ be vectors in im $\mathscr{A}$. This means that in $\mathbf{V}$ there exist vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ such that $\mathscr{A}\mathbf{x}_1 = \mathbf{y}_1$ and $\mathscr{A}\mathbf{x}_2 = \mathbf{y}_2$.

From the formula

$$\lambda_1\mathbf{y}_1 + \lambda_2\mathbf{y}_2 = \lambda_1\mathscr{A}\mathbf{x}_1 + \lambda_2\mathscr{A}\mathbf{x}_2 = \mathscr{A}(\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2)$$

it follows that any linear combination of the vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ also belongs to im $\mathscr{A}$. ▶

The *dimension* of the image of the linear mapping is called the *rank* of the linear mapping.

In symbols, we write rank ($\mathscr{A}$) to mean the rank of the linear mapping $\mathscr{A}$.

The linear mappings $\mathscr{A}$: $\mathbf{V} \to \mathbf{W}$ and $\mathscr{B}$: $\mathbf{V} \to \mathbf{W}$ are said to *coincide* if for any vector $\mathbf{x}$ in $\mathbf{V}$ there holds $\mathscr{A}\mathbf{x} = \mathscr{B}\mathbf{x}$.

In symbols, we write $\mathscr{A} = \mathscr{B}$ to mean the coincidence of $\mathscr{A}$ and $\mathscr{B}$.

Theorem 6.11 (on construction of a linear mapping). *Let* $\mathbf{V}$ *and* $\mathbf{W}$ *be two linear spaces and* $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ *be a basis in* $\mathbf{V}$, *and let* $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n$ *be arbitrary vectors in* $\mathbf{W}$.

Then there exists the only linear mapping

$$\mathscr{A}: \mathbf{V} \to \mathbf{W}$$

such that

$$\mathscr{A}\mathbf{e}_k = \mathbf{f}_k \quad (k = 1, 2, \ldots, n). \tag{6.13}$$

First, we prove that a linear mapping satisfying (6.13) exists.

◀ Consider the expansion of some vector $\mathbf{x}$ in $\mathbf{V}$ with respect to the basis $\mathbf{e}$

$$\mathbf{x} = \sum_{k=1}^{n} \xi^k \mathbf{e}_k$$

and the mapping $\mathscr{A}$: $\mathbf{V} \to \mathbf{W}$ such that

$$\mathscr{A}\mathbf{x} = \sum_{k=1}^{n} \xi^k \mathbf{f}_k. \tag{6.14}$$

It is easy to verify directly that the induced mapping is linear. Let

$$\mathbf{y} = \sum_{k=1}^{n} \eta^k \mathbf{e}_k.$$

Then applying (6.14) we have

$$\mathscr{A}(\lambda\mathbf{x} + \mu\mathbf{y}) = \sum_{k=1}^{n} (\lambda\xi^k + \mu\eta^k)\mathbf{f}_k$$

$$= \lambda \sum_{k=1}^{n} \xi^k\mathbf{f}_k + \mu \sum_{k=1}^{n} \eta^k\mathbf{f}_k = \lambda\,\mathscr{A}\mathbf{x} + \mu\,\mathscr{A}\mathbf{y}.$$

Whence we infer that

$$\mathscr{A}\mathbf{e}_k = \mathbf{f}_k \quad (k = 1, 2, \ldots, n).$$ ►

Now we show that the linear mapping satisfying (6.13) is unique.

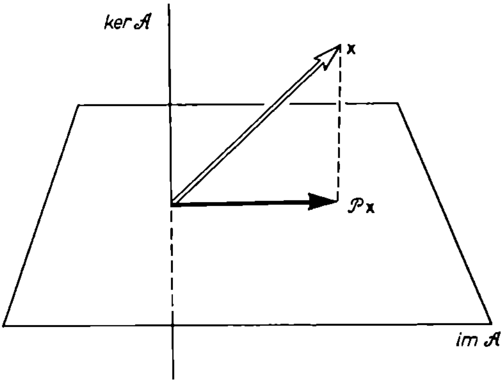


Fig. 6.14

◄ Consider the mapping $\mathscr{A}$ satisfying (6.13) and the mapping $\mathscr{B}: \mathbf{V} \to \mathbf{W}$ such that

$$\mathscr{B}\mathbf{e}_k = \mathbf{f}_k \quad (k = 1, 2, \ldots, n).$$

Computing the actions of $\mathscr{A}$ and $\mathscr{B}$ on an arbitrary vector $\mathbf{x}$ in $\mathbf{V}$, we arrive at the identity

$$\mathscr{A}\mathbf{x} = \sum_{k=1}^{n} \xi^k\mathbf{f}_k = \sum_{k=1}^{n} \xi^k\,\mathscr{B}\mathbf{e}_k = \mathscr{B}\left(\sum_{k=1}^{n} \xi^k\mathbf{e}_k\right) = \mathscr{B}\mathbf{x}.$$

Whence we conclude that the linear mappings $\mathscr{A}$ and $\mathscr{B}$ coincide. ►

From Theorem 6.11 it follows that we may define a linear mapping by identifying its action on the base vectors of $\mathbf{V}$.

The *kernel* of the linear mapping $\mathscr{A}$: $\mathbf{V} \to \mathbf{W}$ is a set of all vectors of $\mathbf{V}$ which are mapped into the zero vector of $\boldsymbol{\theta}_w$ of $\mathbf{W}$.

We denote the kernel of $\mathscr{A}$ by ker $\mathscr{A}$.

Turning back to the examples (1-5) of linear mappings, we can easily notice that

(1″) The kernel of the differential mapping $\mathscr{D}$: $\mathbf{M}_n \to \mathbf{M}_n$ is a set of polynomials of degree zero.

(2″) The kernel of the similarity mapping consists of $\boldsymbol{\theta}_v$ only.

(3″) The kernel of the projective mapping $\mathscr{P}$: $\mathbf{V} \to \mathbf{V}$ is the linear subspace $\mathbf{L}(\mathbf{e}_{k+1}, \mathbf{e}_{k+2}, \ldots, \mathbf{e}_n)$ (Fig. 6.14).

(4″) The kernel of the differential mapping $\mathscr{D}$: $\mathbf{T}_2 \to \mathbf{T}_2$ consists of $\boldsymbol{\theta}_v$ only.

(5″) The set of solutions of the homogeneous linear system

$$\mathbf{AX} = \mathbf{0}$$

form the kernel of the mapping

$$\mathscr{A}: \mathbb{R}^{n \times 1} \to \mathbb{R}^{m \times 1}.$$

Proposition. The kernel of the linear mapping $\mathscr{A}$: $\mathbf{V} \to \mathbf{W}$ is a linear subspace of $\mathbf{V}$.

◄ The identities $\mathscr{A}\mathbf{x} = \boldsymbol{\theta}_w$ and $\mathscr{A}\mathbf{y} = \boldsymbol{\theta}_w$ imply that

$$\mathscr{A}(\lambda\mathbf{x} + \mu\mathbf{y}) = \lambda\mathscr{A}\mathbf{x} + \mu\mathscr{A}\mathbf{y} = \lambda\boldsymbol{\theta}_w + \mu\boldsymbol{\theta}_w = \boldsymbol{\theta}_w \ ►.$$

The *dimension of the kernel* of the linear mapping $\mathscr{A}$ is called the *nullity* of $\mathscr{A}$.

In symbols, we write dim ker $\mathscr{A}$ = nullity $\mathscr{A}$.

Notice that for any linear space $\mathscr{A}$: $\mathbf{V} \to \mathbf{W}$ the following identity holds

$$\text{rank}\,\mathscr{A} + \text{nullity}\,\mathscr{A} = \dim \mathbf{V}. \qquad (*)$$

Operations on linear mappings. Let $\mathscr{A}$: $\mathbf{V} \to \mathbf{W}$ and $\mathscr{B}$: $\mathbf{V} \to \mathbf{W}$ be two linear mappings.

The *sum* of the linear mappings $\mathscr{A}$ and $\mathscr{B}$ is the mapping $\mathscr{C}$: $\mathbf{V} \to \mathbf{W}$ such that given any $\mathbf{x}$ in $\mathbf{V}$

$$\mathscr{C}\mathbf{x} = \mathscr{A}\mathbf{x} + \mathscr{B}\mathbf{x}.$$

It is easy to verify that the mapping $\mathscr{C}$ is linear. Indeed,

$$\begin{aligned}\mathscr{C}(\lambda\mathbf{x} + \mu\mathbf{y}) &= \mathscr{A}(\lambda\mathbf{x} + \mu\mathbf{y}) + \mathscr{B}(\lambda\mathbf{x} + \mu\mathbf{y}) \\ &= \lambda(\mathscr{A}\mathbf{x} + \mathscr{B}\mathbf{x}) + \mu(\mathscr{A}\mathbf{y} + \mathscr{B}\mathbf{y}) = \lambda\mathscr{C}\mathbf{x} + \mu\mathscr{C}\mathbf{y}.\end{aligned}$$

In symbols, we denote the sum of $\mathscr{A}$ and $\mathscr{B}$ as $\mathscr{C} = \mathscr{A} + \mathscr{B}$.

The *product* of the linear mapping $\mathscr{A}$: $\mathbf{V} \to \mathbf{W}$ by an arbitrary number α is a mapping $\mathscr{B}$: $\mathbf{V} \to \mathbf{W}$ such that given any $\mathbf{x}$ in $\mathbf{V}$

$$\mathscr{B}\mathbf{x} = \alpha\,\mathscr{A}\mathbf{x}.$$

The mapping $\mathscr{B}$ is linear since

$$\begin{aligned}\mathscr{B}(\lambda\mathbf{x} + \mu\mathbf{y}) &= \alpha\mathscr{A}(\lambda\mathbf{x} + \mu\mathbf{y}) \\ &= \lambda(\alpha\mathscr{A}\mathbf{x}) + \mu(\alpha\mathscr{A}\mathbf{y}) = \lambda\mathscr{B}\mathbf{x} + \mu\mathscr{B}\mathbf{y}.\end{aligned}$$

In symbols, we write $\mathscr{B} = \alpha\mathscr{A}$.

6.11 Linear Operators

We shall confine ourselves to linear operators, that is, to linear mappings from a space **V** to **V**. Notice that the differential mappings, the identity and projective mappings and pre-multiplication of a square matrix by a column-vector that we have encountered in the previous sections are all examples of linear operators.

Multiplication of linear operators. Let $\mathscr{A}: \mathbf{V} \to \mathbf{V}$ and $\mathscr{B}: \mathbf{V} \to \mathbf{V}$ be two linear operators.

The *product* of the linear operator $\mathscr{A}$ by the linear operator $\mathscr{B}$ is the mapping $\mathscr{C}: \mathbf{V} \to \mathbf{V}$ such that

$$\mathscr{C}\mathbf{x} = \mathscr{B}(\mathscr{A}\mathbf{x}).$$

We show that $\mathscr{C}$ is a linear operator. Indeed

$$\begin{aligned}\mathscr{C}(\lambda\mathbf{x} + \mu\mathbf{y}) &= \mathscr{B}(\mathscr{A}(\lambda\mathbf{x} + \mu\mathbf{y})) = \mathscr{B}(\lambda\mathscr{A}\mathbf{x} + \mu\mathscr{A}\mathbf{y}) \\ &= \lambda\mathscr{B}(\mathscr{A}\mathbf{x}) + \mu\mathscr{B}(\mathscr{A}\mathbf{y}) = \lambda\mathscr{C}\mathbf{x} + \mu\mathscr{C}\mathbf{y}.\end{aligned}$$

In symbols, we write $\mathscr{C} = \mathscr{B}\mathscr{A}$.

Remark. Notice that in general $\mathscr{B}\mathscr{A} \neq \mathscr{A}\mathscr{B}$ as is easily seen from the following example.

Example. Let $\mathbf{V} = \mathbb{R}^2$ (Fig. 6.15).

◄ The mappings

$$\mathscr{A}: (\xi^1, \xi^2) \to (\xi^1, 0) \quad \text{and} \quad \mathscr{B}: (\xi^1, \xi^2) \to (\xi^1 + \xi^2, \xi^2)$$

are linear operators from **V** into **V**.

Then

$$\mathscr{B}\mathscr{A}: (\xi^1, \xi^2) \to (\xi^1, 0) \quad \text{and} \quad \mathscr{A}\mathscr{B}: (\xi^1, \xi^2) \to (\xi^1 + \xi^2, 0).$$

Whence we deduce that

$$\mathscr{B}\mathscr{A}(\xi^1, \xi^2) \neq (\mathscr{A}\mathscr{B})(\xi^1, \xi^2)$$

if $\xi^2 \neq 0$. ►

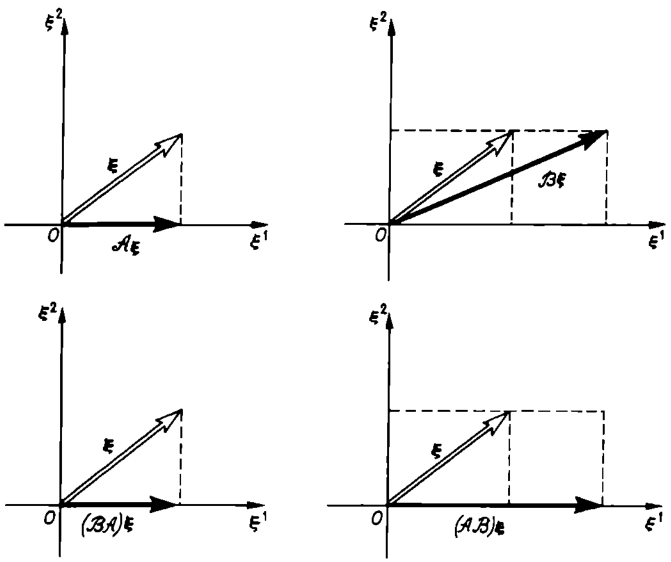


Fig. 6.15

Definition. The linear operators $\mathscr{A}: \mathbf{V} \to \mathbf{V}$ and $\mathscr{B}: \mathbf{V} \to \mathbf{V}$ are said to be *equivalent* if given any vector $\mathbf{x}$ in $\mathbf{V}$ there holds $\mathscr{A}\mathbf{x} = \mathscr{B}\mathbf{x}$.

Let $\mathscr{A}: \mathbf{V} \to \mathbf{V}$ be a linear operator.

The linear operator $\mathscr{B}: \mathbf{V} \to \mathbf{V}$ is said to be an *inverse* of $\mathscr{A}$ if

$$\mathscr{B}\mathscr{A} = \mathscr{A}\mathscr{B} = \mathscr{I}, \tag{6.15}$$

where $\mathscr{I}: \mathbf{V} \to \mathbf{V}$ is the identity operator, i.e., $\mathscr{I}\mathbf{x} = \mathbf{x}$ for any vector $\mathbf{x}$ in $\mathbf{V}$.

Theorem 6.12. *For the linear operator $\mathscr{A}: \mathbf{V} \to \mathbf{V}$ to be invertible it is necessary and sufficient that the image of $\mathscr{A}$ coincide with the space $\mathbf{V}$, i.e.,* $\operatorname{im} \mathscr{A} = \mathbf{V}$.

◄ Suppose, first, that there exists a linear operator $\mathscr{B}$ inverse to the given linear operator $\mathscr{A}$.

Recall that $\operatorname{im} \mathscr{A}$ of $\mathscr{A}$ is a subspace of $\mathbf{V}$.

We show that an arbitrary vector $\mathbf{y}$ in $\mathbf{V}$ belongs to $\operatorname{im} \mathscr{A}$.

Let $\mathbf{x} = \mathscr{B}\mathbf{y}$. By (6.15) we have

$$\mathscr{A}\mathbf{x} = \mathscr{A}(\mathscr{B}\mathbf{y}) = (\mathscr{A}\mathscr{B})\mathbf{y} = \mathscr{I}\mathbf{y} = \mathbf{y}.$$

Whence we infer that the vector $\mathbf{y}$ is the image of the vector $\mathbf{x} = \mathscr{B}\mathbf{y}$ and, consequently, $\mathbf{y}$ belongs to $\operatorname{im} \mathscr{A}$. Therefore, $\operatorname{im} \mathscr{A} = \mathbf{V}$. ►

Now we suppose that the image of $\mathscr{A}$ coincides with the space $\mathbf{V}$, i.e., $\operatorname{im} \mathscr{A} = \mathbf{V}$. Then

$$\operatorname{rank} \mathscr{A} = \dim \mathbf{V}$$

and the linear operator $\mathscr{A}$ maps a basis of $\mathbf{V}$ into another basis of $\mathbf{V}$, i.e.,

$$\mathscr{A}\colon \mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n) \to \mathbf{f} = (\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n),$$

where $\mathbf{f}_k = \mathscr{A}\mathbf{e}_k$ $(k = 1, 2, \ldots, n)$.

Consider a linear operator $\mathscr{B}$ such that

$$\mathscr{B}\mathbf{f}_k = \mathbf{e}_k \quad (k = 1, 2, \ldots, n). \tag{6.16}$$

Theorem 6.11 states that the linear operator $\mathscr{B}$ satisfying (6.16) is unique.

Let us compute $(\mathscr{B}\mathscr{A})\mathbf{x}$ and $(\mathscr{A}\mathscr{B})\mathbf{x}$ for an arbitrary vector $\mathbf{x}$ in $\mathbf{V}$.

◄ Relative to the basis $\mathbf{e}$ the vector $\mathbf{x}$ takes the form

$$\mathbf{x} = \sum_{k=1}^{n} \xi^k \mathbf{e}_k.$$

Using (6.16), we obtain

$$\begin{aligned}(\mathscr{B}\mathscr{A})\mathbf{x} &= (\mathscr{B}\mathscr{A})\left(\sum_{k=1}^{n} \xi^k \mathbf{e}_k\right) = \mathscr{B}\left(\sum_{k=1}^{n} \xi^k \mathscr{A}\mathbf{e}_k\right) \\ &= \sum_{k=1}^{n} \xi^k\, \mathscr{B}(\mathscr{A}\mathbf{e}_k) = \sum_{k=1}^{n} \xi^k \mathscr{B}\mathbf{f}_k = \sum_{k=1}^{n} \xi^k \mathbf{e}_k = \mathbf{x}.\end{aligned}$$

Analogously, relative to the basis $\mathbf{f}$ we have

$$\mathbf{x} = \sum_{k=1}^{n} \eta^k \mathbf{f}_k$$

and

$$(\mathscr{A}\mathscr{B})\mathbf{x} = \sum_{k=1}^{n} \eta^k\, \mathscr{A}(\mathscr{B}\mathbf{f}_k) = \sum_{k=1}^{n} \eta^k\, \mathscr{A}\mathbf{e}_k = \sum_{k=1}^{n} \eta^k \mathbf{f}_k = \mathbf{x}.$$

Thus for any vector $\mathbf{x}$ in $\mathbf{V}$ we have

$$\mathscr{B}\mathscr{A}\mathbf{x} = \mathbf{x} \quad \text{and} \quad \mathscr{A}\mathscr{B}\mathbf{x} = \mathbf{x}.$$

Whence we conclude that

$$\mathscr{B}\mathscr{A} = \mathscr{A}\mathscr{B} = \mathscr{I}$$ ►

Remark. Theorem 6.12 states that the inverse $\mathscr{B}$ of $\mathscr{A}$ is uniquely defined.

In symbols, we write $\mathscr{A}^{-1}$ to mean the inverse of the linear operator $\mathscr{A}$.

Corollary. The linear operator $\mathscr{A}\colon \mathbf{V} \to \mathbf{V}$ is invertible if and only if the kernel of $\mathscr{A}$ consists of the zero vector only, i.e.,

$$\ker \mathscr{A} = \{\boldsymbol{\theta}_V\}.$$

This immediately follows from Theorem 6.12 and formula (*) on page 196.

Example. The linear operator

$$\mathscr{A}: (\xi^1, \xi^2) \to \left(\xi^1, \frac{2}{3}\xi^2\right)$$

compresses the plane uniformly (with coefficient 2/3) towards the axis ξ^1 while the inverse operator

$$\mathscr{A}^{-1}: (\xi^1, \xi^2) \to \left(\xi^1, \frac{3}{2}\xi^2\right)$$

extends the plane uniformly with coefficient 3/2 (Fig. 6.16).

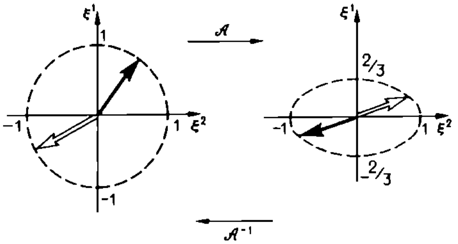


Fig. 6.16

6.12 Matrices of Linear Operators

Suppose that the linear operator $\mathscr{A}$: $\mathbf{V} \to \mathbf{V}$ acts on the vectors $\mathbf{e}_1$, $\mathbf{e}_2$, ..., $\mathbf{e}_n$ of the basis $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n)$ in $\mathbf{V}$ so that

$$\mathscr{A}\mathbf{e}_i = \sum_{k=1}^{n} \alpha_i^k \mathbf{e}_k = \alpha_i^1\mathbf{e}_1 + \alpha_i^2\mathbf{e}_2 + \ldots + \alpha_i^n\mathbf{e}_n \quad (i = 1, 2, \ldots, n).$$

The matrix

$$\mathbf{A} = \mathbf{A}(\mathbf{e}) = \begin{pmatrix} \alpha_1^1 & \alpha_2^1 & \ldots & \alpha_n^1 \\ \alpha_1^2 & \alpha_2^2 & \ldots & \alpha_n^2 \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_1^n & \alpha_2^n & \ldots & \alpha_n^n \end{pmatrix}$$

whose columns comprise the coordinates of the images of the base vectors is called the *matrix of the linear operator* $\mathscr{A}$ relative to the basis **e**.

Examples. (1) Relative to the basis $\mathbf{e}_0 = 1$, $\mathbf{e}_1 = t$, $\mathbf{e}_2 = \frac{t^2}{2!}$, $\mathbf{e}_3 = \frac{t^3}{3!}$ the matrix $\mathbf{D}(\mathbf{e})$ of the differential linear operator $\mathscr{D}$: $\mathbf{M}_3 \to \mathbf{M}_3$ takes the form

$$\mathbf{D}(\mathbf{e}) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

(2) Relative to the basis $\mathbf{e}_1 = \cos x$, $\mathbf{e}_2 = \sin x$ the matrix $\mathbf{D}(\mathbf{e})$ of the differential linear operator $\mathscr{D}\colon \mathbf{T}_2 \to \mathbf{T}_2$ takes the form

$$\mathbf{D}(\mathbf{e}) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

since

$$\mathscr{D}\mathbf{e}_1 = \mathscr{D}\cos x = -\sin x = -\mathbf{e}_2 = 0\cdot\mathbf{e}_1 + (-1)\cdot\mathbf{e}_2$$

and

$$\mathscr{D}\mathbf{e}_2 = \mathscr{D}\sin x = \cos x = \mathbf{e}_1 = 1\cdot\mathbf{e}_1 + 0\cdot\mathbf{e}_2. \blacktriangleright$$

Let

$$\mathbf{y} = \mathscr{A}\mathbf{x}$$

and let

$$\mathbf{x} = \sum_{i=1}^{n} \xi^i \mathbf{e}_i \quad \text{and} \quad \mathbf{y} = \sum_{k=1}^{n} \eta^k \mathbf{e}_k$$

be the expansions of the vectors $\mathbf{x}$ and $\mathbf{y}$ relative to the basis $\mathbf{e}$.

Then the column-vectors

$$\mathbf{x}(\mathbf{e}) = \begin{pmatrix} \xi^1 \\ \xi^2 \\ \vdots \\ \xi^n \end{pmatrix} \quad \text{and} \quad \mathbf{y}(\mathbf{e}) = \begin{pmatrix} \eta^1 \\ \eta^2 \\ \vdots \\ \eta^n \end{pmatrix}$$

which comprise the coordinates of $\mathbf{x}$ and $\mathbf{y}$ relative to the basis $\mathbf{e}$ are related as

$$\mathbf{y}(\mathbf{e}) = \mathbf{A}(\mathbf{e})\mathbf{x}(\mathbf{e}). \tag{6.17}$$

◀ Indeed, since the expansion of the vector $\mathbf{y}$ relative to the basis $\mathbf{e}$ is unique, comparing

$$\mathbf{y} = \sum_{k=1}^{n} \eta^k \mathbf{e}_k$$

with

$$\mathbf{y} = \mathscr{A}\mathbf{x} = \sum_{i=1}^{n} \xi^i \mathscr{A}\mathbf{e}_i = \sum_{i=1}^{n} \xi^i \left(\sum_{k=1}^{n} \alpha_i^k \mathbf{e}_k \right)$$

$$= \sum_{k=1}^{n} \left(\sum_{i=1}^{n} \alpha_i^k \xi^i \right) \mathbf{e}_k$$

yields

$$\eta^k = \sum_{i=1}^{n} \alpha_i^k \xi^i \quad (k = 1, 2, \dots, n).$$

Arranging these n identities in matrix form, we have

$$\begin{pmatrix} \eta^1 \\ \eta^2 \\ \vdots \\ \eta^n \end{pmatrix} = \begin{pmatrix} \alpha_1^1 & \alpha_2^1 & \dots & \alpha_n^1 \\ \alpha_1^2 & \alpha_2^2 & \dots & \alpha_n^2 \\ \dots & \dots & \dots & \dots \\ \alpha_1^n & \alpha_2^n & \dots & \alpha_n^n \end{pmatrix} \begin{pmatrix} \xi^1 \\ \xi^2 \\ \vdots \\ \xi^n \end{pmatrix}.$$

Whence follows (6.17). ►

Theorem 6.13. *The rank of the matrix* **A**(**e**) *of the linear operator* $\mathscr{A}$: **V** → **V** *is independent of the choice of the basis* **e** *and is equal to the rank of* $\mathscr{A}$.

◄ Since

$$\operatorname{im} \mathscr{A} = \mathbf{L}(\mathscr{A}\mathbf{e}_1, \mathscr{A}\mathbf{e}_2, \dots, \mathscr{A}\mathbf{e}_n)$$

the rank of $\mathscr{A}$ is equal to the maximal number of linearly independent vectors belonging to the collection of the vectors $\mathscr{A}\mathbf{e}_1, \mathscr{A}\mathbf{e}_2, \dots, \mathscr{A}\mathbf{e}_n$. By virtue of Theorem 6.4 this number is equal to the maximal number of linearly independent column-vectors in the matrix **A**(**e**), i.e to the rank of **A**(**e**).

Therefore

$$r(\mathbf{A}(\mathbf{e})) = \operatorname{rank} \mathscr{A}. \ ►$$

It is easy to verify that addition and multiplication of the linear operators become addition and multiplication of the matrices of these operators relative to the same basis and multiplication of a linear operator by a scalar becomes multiplication of the matrix of the operator by this scalar.

◄ By way of illustration we show that the matrix of $\mathscr{C} = \mathscr{B}\mathscr{A}$ is equal to the product of the matrices **A**(**e**) and **B**(**e**) of the linear operators $\mathscr{A}$ and $\mathscr{B}$ relative to the basis **e**, i.e.,

$$\mathbf{C}(\mathbf{e}) = \mathbf{A}(\mathbf{e})\mathbf{B}(\mathbf{e}).$$

Let

$$\mathscr{A}\mathbf{e}_i = \sum_{k=1}^{n} \alpha_i^k \mathbf{e}_k \quad \text{and} \quad \mathscr{B}\mathbf{e}_k = \sum_{m=1}^{n} \beta_k^m \mathbf{e}_m.$$

Then

$$\mathscr{C}\mathbf{e}_i = \mathscr{B}\mathscr{A}\mathbf{e}_i = \mathscr{B}\left(\sum_{k=1}^{n} \alpha_i^k \mathbf{e}_k\right) = \sum_{k=1}^{n} \alpha_i^k \mathscr{B}\mathbf{e}_k$$

$$= \sum_{k=1}^{n} \alpha_i^k \left(\sum_{m=1}^{n} \beta_k^m \mathbf{e}_k\right) = \sum_{m=1}^{n} \left(\sum_{k=1}^{n} \beta_k^m \alpha_i^k\right) \mathbf{e}_k.$$

Setting

$$\gamma_i^m = \sum_{k=1}^{n} \beta_k^m \alpha_i^k \quad (i,\ m = 1,\ 2,\ \dots,\ n) \tag{6.18}$$

we may write

$$\mathbf{C}(\mathbf{e}) = (\gamma_i^m). \tag{6.19}$$

On the other hand, since $\mathbf{A}(\mathbf{e}) = (\alpha_i^k)$ and $\mathbf{B}(\mathbf{e}) = (\beta_k^m)$ (6.18) and (6.19) yield

$$\mathbf{C}(\mathbf{e}) = \mathbf{B}(\mathbf{e})\mathbf{A}(\mathbf{e}). \tag{6.20}$$

Thus, relative to the basis $\mathbf{e}$, the matrix of the linear operator $\mathscr{B}\mathscr{A}$ is equal to $\mathbf{B}(\mathbf{e})\mathbf{A}(\mathbf{e})$. ►

From the equivalence of the operation of multiplication of linear operators and that of multiplication of the respective matrices it easily follows that the matrix of the linear operator $\mathscr{A}^{-1}$ which is inverse to the operator $\mathscr{A}$ is the inverse of the matrix $\mathbf{A}$ of $\mathscr{A}$.

◄ Indeed, by definition of the inverse operator we have

$$\mathscr{A}^{-1}\mathscr{A} = \mathscr{I} \quad \text{and} \quad \mathscr{A}\mathscr{A}^{-1} = \mathscr{I}.$$

This means that the matrix of $\mathscr{A}^{-1}$, say $\mathbf{B}$, must satisfy the identities

$$\mathbf{BA} = \mathbf{I} \quad \text{and} \quad \mathbf{AB} = \mathbf{I}.$$

Whence we conclude that $\mathbf{B}$ is the inverse of $\mathbf{A}$, i.e.,

$$\mathbf{B} = \mathbf{A}^{-1}.$$ ►

Theorem 6.14. *Let* $\mathscr{A}\colon \mathbf{V} \to \mathbf{V}$ *be a linear operator and let* $\mathbf{A} = \mathbf{A}(\mathbf{e})$ *and* $\mathbf{A}' = \mathbf{A}(\mathbf{e}')$ *be matrices of* $\mathscr{A}$ *relative to the bases* $\mathbf{e}$ *and* $\mathbf{e}'$ *of a linear space* $\mathbf{V}$.

Then

$$\mathbf{A}' = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}, \tag{6.21}$$

where $\mathbf{S}$ *is the matrix of transition from* $\mathbf{e}$ *to* $\mathbf{e}'$.

◄ Let $\mathbf{y} = \mathscr{A}\mathbf{x}$.

Relative to the bases $\mathbf{e}$ and $\mathbf{e}'$ the column-vectors of the coordinates of $\mathbf{x}$ and $\mathbf{y}$ are related as

$$\mathbf{y}(\mathbf{e}) = \mathbf{A}\mathbf{x}(\mathbf{e}) \quad \text{and} \quad \mathbf{y}(\mathbf{e}') = \mathbf{A}'\mathbf{x}(\mathbf{e}'). \tag{6.22}$$

By Property (b) of transition matrices we may write

$$\mathbf{x}(\mathbf{e}) = \mathbf{S}\mathbf{x}(\mathbf{e}') \quad \text{and} \quad \mathbf{y}(\mathbf{e}) = \mathbf{S}\mathbf{y}(\mathbf{e}'). \tag{6.23}$$

Substituting (6.23) into the first identity of (6.22), we obtain

$$\mathbf{S}\mathbf{y}(\mathbf{e}') = \mathbf{A}\mathbf{S}\mathbf{x}(\mathbf{e}').$$

Whence, using the second identity of (6.22), we have

$$\mathbf{SA'x(e') = ASx(e')}.$$

Notice that the above identity holds for any vector $\mathbf{x}$. Hence we may write

$$\mathbf{SA' = AS}.$$

Since the transition matrix $\mathbf{S}$ is nonsingular and, consequently, invertible, pre-multiplication of the last identity by $\mathbf{S}^{-1}$ gives (6.21). ►

Corollary. The determinant of the matrix of a linear operator remains unchanged in any basis.

◄ Let us compute the determinant of the matrix

$$\mathbf{A(e') = S^{-1}A(e)S}.$$

We have

$$\det \mathbf{A(e')} = \det(\mathbf{S^{-1}A(e)S}) = \det \mathbf{S}^{-1}\cdot\det \mathbf{A(e)}\cdot\det \mathbf{S} = \det \mathbf{A(e)},$$

since

$$\det \mathbf{S}^{-1} = (\det \mathbf{S})^{-1}. \ ►$$

It is easy to verify that the determinant of the matrix of the linear operator

$$\mathscr{A} - t\mathscr{I},$$

where $\mathscr{I}$ is the identity operator and t is an arbitrary number, also remains unchanged in any basis.

◄ Indeed, let $\mathbf{A(e)} - t\mathbf{I}$ and $\mathbf{A(e')} - t\mathbf{I}$ be matrices of the operator $\mathscr{A} - t\mathscr{I}$ relative to the bases $\mathbf{e}$ and $\mathbf{e'}$.

Using (6.21), we obtain

$$\mathbf{A(e')} - t\mathbf{I} = \mathbf{S^{-1}A(e)S} - t\mathbf{I} = \mathbf{S}^{-1}(\mathbf{A(e)} - t\mathbf{I})\mathbf{S}.$$

Then by Corollary of Theorem 6.14 we have

$$\det(\mathbf{A(e')} - t\mathbf{I}) = \det(\mathbf{A(e)} - t\mathbf{I}). \ ►$$

Expanding the determinant

$$\det(\mathbf{A} - t\mathbf{I}) = \begin{vmatrix} \alpha_1^1 - t & \alpha_2^1 & \dots & \alpha_n^1 \\ \alpha_1^2 & \alpha_2^2 - t & \dots & \alpha_n^2 \\ \dots & \dots & \dots & \dots \\ \alpha_1^n & \alpha_2^n & \dots & \alpha_n^n - t \end{vmatrix}$$

$$= (-1)^n t^n + \gamma_{n-1}t^{n-1} + \dots + \gamma_1 t + \gamma_0,$$

we can write

$$\chi(t) = \det(\mathbf{A} - t\mathbf{I}) = (-1)^n t^n + \gamma_{n-1}t^{n-1} + \dots + \gamma_1 t + \gamma_0.$$

The polynomial $\chi(t)$ is called the *characteristic polynomial of the linear operator* $\mathscr{A}$ and the roots of $\chi(t)$ are called the *eigenvalues of the linear operator* $\mathscr{A}$ (the matrix **A**). Notice that the characteristic polynomial $\chi(t)$ is independent of the choice of a basis.

6.13 Eigenvalues and Eigenvectors

The nonzero vector $\mathbf{x}$ in $\mathbf{V}$ is called the *eigenvector* of the linear operator $\mathscr{A}: \mathbf{V} \to \mathbf{V}$ if there exists the eigenvalue λ of $\mathscr{A}$ such that

$$\mathscr{A}\mathbf{x} = \lambda\mathbf{x}. \tag{$*$}$$

Examples. (1) Any polynomial of the zero degree is the eigenvector of the differential operator

$$\mathscr{D} = \frac{d}{dt}: \quad \mathbf{M}_n \to \mathbf{M}_n.$$

◄ Consider a polynomial of the zero degree, which is equal to 1. Then applying ($*$), we have

$$\frac{d}{dt}(1) = 0 \times 1 = 0.$$

Whence it follows that the eigenvalue λ of $\mathscr{D}$ is equal to zero. ►

(2) The differential operator

$$\mathscr{D} = \frac{d}{dt}: \quad \mathbf{T}_2 \to \mathbf{T}_2$$

has no eignvectors.

◄ Consider the trigonometric polynomial $\alpha \cos t + \beta \sin t$.

Using ($*$), we have

$$\mathscr{D}(\alpha \cos t + \beta \sin t) = \lambda (\alpha \cos t + \beta \sin t).$$

This means that

$$-\alpha \sin t + \beta \cos t = \lambda\alpha \cos t + \lambda\beta \sin t,$$

or

$$(\lambda\beta + \alpha) \sin t + (\lambda\alpha - \beta) \cos t = 0,$$

which is fulfilled if and only if

$$\alpha + \lambda\beta = 0 \quad \text{and} \quad \lambda\alpha - \beta = 0.$$

Whence $\alpha = \beta = 0$. Consequently, the polynomial $\alpha \cos t + \beta \sin t$ is a zero one and can not be the eigenvector of $\mathscr{D}$. ►

Theorem 6.15. *The number* λ *is the eigenvalue of the linear operator* $\mathscr{A}$ *if and only if* λ *is the root of the characteristic polynomial* $\chi(t)$ *of* $\mathscr{A}$ *i.e.* $\chi(\lambda) = 0$.

◄ Suppose that λ is the root of the polynomial $\chi(t)$ in which case

$$\chi(\lambda) = \det(\mathbf{A}(\mathbf{e}) - \lambda\mathbf{I}) = 0. \tag{6.24}$$

Consider the homogeneous linear system

$$(\mathbf{A}(\mathbf{e}) - \lambda\mathbf{I})\begin{pmatrix}\xi^1\\ \xi^2\\ \vdots\\ \xi^n\end{pmatrix} = \begin{pmatrix}0\\0\\ \vdots\\ 0\end{pmatrix}$$

or

$$\begin{array}{llll}(\alpha_1^1 - \lambda)\xi^1 + \alpha_2^1\xi^2 & & + \dots + \alpha_n^1\xi^n & = 0,\\ \alpha_2^2\xi^1 & + (\alpha_2^2 - \lambda)\xi^2 & + \dots + \alpha_n^2\xi^n & = 0,\\ \dots\dots\dots & \dots\dots\dots & \dots\dots\dots & \dots\\ \alpha_1^n\xi^1 & + \alpha_2^n\xi^2 & + \dots + (\alpha_n^n - \lambda)\xi^n & = 0.\end{array} \tag{6.25}$$

By virtue of (6.24) system (6.25) has the nonzero solution $\xi^1, \xi^2, \dots, \xi^n$ and

$$\mathbf{x} = \sum_{i=1}^{n} \xi^i\mathbf{e}_i \neq \boldsymbol{\theta}_v.$$

The column-vector $\mathbf{x}(\mathbf{e})$ of the coordinates of $\mathbf{x}$ satisfies the condition

$$(\mathbf{A}(\mathbf{e}) - \lambda\mathbf{I})\mathbf{x}(\mathbf{e}) = 0,$$

which may be written as

$$\mathbf{A}(\mathbf{e})\mathbf{x}(\mathbf{e}) = \lambda\mathbf{x}(\mathbf{e}),$$

or, equivalently,

$$\mathscr{A}\mathbf{x} = \lambda\mathbf{x}.$$

Whence it follows that λ is the eigenvalue of the linear operator $\mathscr{A}$ and $\mathbf{x}$ is the eigenvector of $\mathscr{A}$.

Now we assume that λ is the eigenvalue of $\mathscr{A}$. Then there exists a nonzero vector $\mathbf{x}$ such that $\mathscr{A}\mathbf{x} = \lambda\mathbf{x}$.

Let $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n)$ be a basis. Then we may write the matrix equation

$$\mathbf{A}(\mathbf{e})\mathbf{x}(\mathbf{e}) = \lambda\mathbf{x}(\mathbf{e}),$$

or

$$(\mathbf{A}(\mathbf{e}) - \lambda\mathbf{I})\mathbf{x}(\mathbf{e}) = 0. \tag{6.26}$$

Since $\mathbf{x}$ is the eigenvector of $\mathscr{A}$ the column-vector $\mathbf{x}(\mathbf{e})$ of the coordinates of $\mathbf{x}$ is a nonzero vector. This means that system (6.26) has a nonzero solution. In this case the condition

$$\det(\mathbf{A}(\mathbf{e}) - \lambda\mathbf{I}) = 0$$

must be fulfilled.

Thus

$$\chi(\lambda) = 0. \blacktriangleright$$

Remark. To find all the eigenvectors corresponding to the given eigenvalue λ it is necessary to construct the fundamental system of solutions for (6.25).

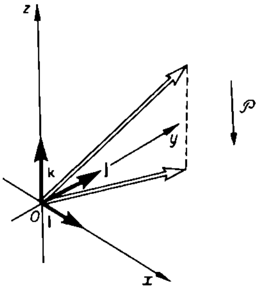


Fig. 6.17

Examples. (1) Let

$$\mathscr{P}\colon \mathbf{V}^3 \to \mathbf{V}^3$$

be a projective linear operator such that

$$\mathscr{P}\colon x\mathbf{i} + y\mathbf{j} + z\mathbf{k} \to x\mathbf{i} + y\mathbf{j}.$$

Compute the eigenvectors of $\mathscr{P}$.

◀ Let us consider the action of $\mathscr{P}$ on the base vectors $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ (Fig. 6.17). We have

$$\mathscr{P}\colon \mathbf{i} \to \mathbf{i},\ \mathscr{P}\colon \mathbf{j} \to \mathbf{j} \text{ and } \mathscr{P}\colon \mathbf{k} \to \boldsymbol{\theta}.$$

Since the matrix of $\mathscr{P}$ is

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

the characteristic polynomial of $\mathscr{P}$ becomes

$$\begin{vmatrix} 1-\lambda & 0 & 0 \\ 0 & 1-\lambda & 0 \\ 0 & 0 & -\lambda \end{vmatrix} = -\lambda(\lambda-1)^2$$

with roots $\lambda_1 = 0$ and λ_2, $\lambda_3 = 1$.

Consider the homogeneous linear systems corresponding to the distinct values of λ equal to 0 and 1. These are defined by the coefficient matrices

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

respectively. Then we may write the homogeneous linear systems as

$$\begin{cases} x \qquad = 0, \\ \quad y \quad = 0, \\ \qquad 0 = 0 \end{cases} \text{ and } \begin{cases} 0 = 0, \\ 0 = 0, \\ -z = 0. \end{cases}$$

The corresponding fundamental systems of solutions are

$$\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

Whence we deduce that the eigenvectors of $\mathscr{P}$ are the base vector **k** and any vector of the form $x\mathbf{i} + y\mathbf{j}$ $(x^2 + y^2 > 0)$ associated with the eigenvalues 0 and 1, respectively. ►

(2) Let

$$\mathscr{D} = \frac{d}{dt} : \alpha + \beta t + \gamma t^2 \to \beta + 2\gamma t$$

be a differential linear operator defined over a space of polynomials of degree not exceeding 2.

Compute the eigenvectors of $\mathscr{D}$.

◄ Relative to the basis 1, t, t^2 the matrix **D** of $\mathscr{D}$ takes the form

$$\mathbf{D} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

Then the characteristic polynomial $-\lambda^3 = 0$ has the root $\lambda = 0$ of multiplicity 3.

Consider the homogeneous linear system

$$\left.\begin{aligned} \beta &= 0 \\ 2\gamma &= 0 \\ 0 &= 0 \end{aligned}\right\}$$

whose solution is 1, 0, 0. The triple 1, 0, 0 corresponds to the polynomial of the zero degree which is the eigenvector of $\mathscr{D}$. ►

6.14 Adjoint Operators

In a finite-dimensional Euclidean space a linear operator can be subjected to an operation which gives rise to a linear operator called the *adjoint operator.*

Let $\mathbf{V}$ be an n-dimensional Euclidean space and $\mathscr{A}: \mathbf{V} \to \mathbf{V}$ be a linear operator.

We shall associate with $\mathscr{A}$ a linear operator $\mathscr{A}^*: \mathbf{V} \to \mathbf{V}$ such that for any vector $\mathbf{x}$ and any vector $\mathbf{y}$ in $\mathbf{V}$ there holds

$$(\mathscr{A}\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathscr{A}^*\mathbf{y}). \tag{6.27}$$

Definition. The linear operator $\mathscr{A}^*: \mathbf{V} \to \mathbf{V}$ satisfying (6.27) is called the *adjoint operator* of the linear operator $\mathscr{A}: \mathbf{V} \to \mathbf{V}$.

Let us show that for any linear operator $\mathscr{A}$ there always exists the adjoint operator $\mathscr{A}^*$.

◄ Let $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ be an orthonormal basis of $\mathbf{V}$ and $\mathbf{A} = \mathbf{A}(\mathbf{e}) = (\alpha_i^j)$ be a matrix of the linear operator $\mathscr{A}$ relative to $\mathbf{e}$, in which case

$$\mathscr{A}\mathbf{e}_i = \sum_{j=1}^{n} \alpha_i^j \mathbf{e}_j \quad (i = 1, 2, \ldots, n). \tag{6.28}$$

Transposing the matrix $\mathbf{A}$ and setting $\mathbf{B} = \mathbf{A}'$, i.e., $\beta_j^i = \alpha_i^j$ for any i and any j $(i, j = 1, 2, \ldots, n)$, we may define the linear operator $\mathscr{A}^*: \mathbf{V} \to \mathbf{V}$ such that

$$\mathscr{A}^*\mathbf{e}_i = \sum_{j=1}^{n} \beta_i^j \mathbf{e}_j \quad (i = 1, 2, \ldots, n), \tag{6.29}$$

where

$$\beta_i^j = \alpha_j^i \text{ and } i, j = 1, 2, \ldots, n. \tag{6.30}$$

(Recall that by virtue of Theorem 6.11 a linear operator is uniquely defined by its action on the base vectors of $\mathbf{V}$.)

To verify (6.27) we first set $\mathbf{x}$ and $\mathbf{y}$ equal to the base vectors, say, $\mathbf{x} = \mathbf{e}_i$ and $\mathbf{y} = \mathbf{e}_j$.

Since the basis $\mathbf{e}$ is orthonormal (6.28) gives

$$(\mathscr{A}\mathbf{e}_i, \mathbf{e}_j) = \left(\sum_{k=1}^{n} \alpha_i^k \mathbf{e}_k, \mathbf{e}_j\right) = \sum_{k=1}^{n} \alpha_i^k (\mathbf{e}_k, \mathbf{e}_j) = \sum_{k=1}^{n} \alpha_i^k \delta_{kj} = \alpha_i^j.$$

Analogously, using (6.29) and (6.30), we obtain

$$(\mathbf{e}_i, \mathscr{A}^*\mathbf{e}_j) = \left(\mathbf{e}_i, \sum_{k=1}^{n} \beta_j^k \mathbf{e}_k\right) = \sum_{k=1}^{n} \beta_j^k (\mathbf{e}_i, \mathbf{e}_k) = \sum_{k=1}^{n} \beta_j^k \delta_{ik} = \beta_j^i = \alpha_i^j.$$

Whence

$$(\mathscr{A}\mathbf{e}_i, \mathbf{e}_j) = (\mathbf{e}_i, \mathscr{A}^*\mathbf{e}_j) \quad (i, j = 1, 2, \ldots, n). \tag{6.31}$$

Now we suppose that $\mathbf{x}$ and $\mathbf{y}$ are arbitrary vectors in $\mathbf{V}$. Relative to the orthonormal basis $\mathbf{e}$ the expansions of $\mathbf{x}$ and $\mathbf{y}$ are

$$\mathbf{x} = \sum_{i=1}^{n} \xi^i \mathbf{e}_i \text{ and } \mathbf{y} = \sum_{j=1}^{n} \eta^j \mathbf{e}_j.$$

Then computing the expressions on the left- and right-hand sides of (6.27), we have

$$(\mathscr{A}\mathbf{x}, \mathbf{y}) = \left(\mathscr{A}\left(\sum_{i=1}^{n} \xi^i \mathbf{e}_i\right), \sum_{j=1}^{n} \eta^j \mathbf{e}_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \xi^i \eta^j (\mathscr{A}\mathbf{e}_i, \mathbf{e}_j)$$

and

$$(\mathbf{x}, \mathscr{A}^*\mathbf{y}) = \left(\sum_{i=1}^{n} \xi^i \mathbf{e}_i, \mathscr{A}^*\left(\sum_{j=1}^{n} \eta^j \mathbf{e}_j\right)\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \xi^i \eta^j (\mathbf{e}_i, \mathscr{A}^*\mathbf{e}_j).$$

Whence by virtue of (6.31) we arrive at the desired result. ►

Example. Consider a two-dimensional linear space of polynomials of degree not exceeding 1 with real coefficients.

Let

$$\varphi(t) = a + bt \text{ and } \psi(t) = c + dt$$

be any two polynomials of degree not exceeding 1 and let

$$(\varphi, \psi) = ac + bd \qquad (*)$$

be a scalar product of φ and ψ.

Then a two-dimensional space of polynomials of degree not exceeding 1 with real coefficients becomes a two-dimensional Euclidean space $\mathbf{M}_1$.

Let $\mathscr{D}: \mathbf{M}_1 \to \mathbf{M}_1$ be a differential operator such that $\mathscr{D}(a + bt) = b$.

We can define the adjoint operator $\mathscr{D}^*: \mathbf{M}_1 \to \mathbf{M}_1$ as follows.

Notice that the polynomials 1 and t form the orthonormal basis of $\mathbf{M}_1$ since using (*), we obtain $(1, 1) = (t, t) = 1$ and $(1, t) = 0$, in which case $\mathscr{D}(1) = 0$ and $\mathscr{D}(t) = 1$, and the matrix of $\mathscr{D}$ becomes $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$.

Then $\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ is the matrix of the adjoint operator $\mathscr{D}^*$ such that $\mathscr{D}^*(1) = t$ and $\mathscr{D}^*(t) = 0$.

For an arbitrary polynomial $\varphi(t) = a + bt$ we have $\mathscr{D}: a + bt \to b$ and $\mathscr{D}^*: a + bt \to at$.

Properties of adjoint operators. (a) Every linear operator has a single adjoint operator.

◄ Suppose that $\mathscr{B}$ and $\mathscr{C}$ are two distinct adjoint operators of a given linear operator $\mathscr{A}$. Then for any $\mathbf{x}$ and any $\mathbf{y}$ in $\mathbf{V}$ there hold

$$(\mathscr{A}\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathscr{B}\mathbf{y}) \quad \text{and} \quad (\mathscr{A}\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathscr{C}\mathbf{y}).$$

Whence

$$(\mathbf{x}, \mathscr{B}\mathbf{y}) = (\mathbf{x}, \mathscr{C}\mathbf{y})$$

and

$$(\mathbf{x}, \mathscr{B}\mathbf{y} - \mathscr{C}\mathbf{y}) = 0.$$

Since **x** is an arbitrary vector in **V** we infer that the vector $\mathscr{B}\mathbf{y} - \mathscr{C}\mathbf{y}$ is orthogonal to every vector in **V** and, consequently, to itself. The latter implies that $\mathscr{B}\mathbf{y} - \mathscr{C}\mathbf{y} = \boldsymbol{\theta}$ and $\mathscr{B}\mathbf{y} = \mathscr{C}\mathbf{y}$. Hence, $\mathscr{B} = \mathscr{C}$ since **y** is an arbitrary vector in **V**. ►

From Property (a) it immediately follows that

(b) $(\alpha\mathscr{A})^* = \alpha\mathscr{A}^*$, where α is an arbitrary real number;

(c) $(\mathscr{A} + \mathscr{B})^* = \mathscr{A}^* + \mathscr{B}^*$;

(d) $(\mathscr{A}\mathscr{B})^* = \mathscr{B}^*\mathscr{A}^*$;

(e) $(\mathscr{A}^*)^* = \mathscr{A}$.

We shall also mention the other two important properties of adjoint operators, namely

(f) Let e be an orthonormal basis of **V**. Then for the linear operators $\mathscr{A}$: $\mathbf{V} \to \mathbf{V}$ and $\mathscr{B}$: $\mathbf{V} \to \mathbf{V}$ to be mutually adjoint, i.e., to satisfy both $\mathscr{B} = \mathscr{A}^*$ and $\mathscr{A} = \mathscr{B}^*$ it is necessary and sufficient that relative to the basis **e** the matrix of one of the operators, say $\mathbf{B} = \mathbf{B}(\mathbf{e})$, be obtained by transposing the matrix $\mathbf{A} = \mathbf{A}(\mathbf{e})$ of the other operator so that $\mathbf{B} = \mathbf{A}'$.

Notice that this property holds true only if **A** and **B** are matrices arranged relative to an orthonormal basis and violates otherwise.

(g) If a linear operator $\mathscr{A}$ is nonsingular so is the adjoint operator $\mathscr{A}^*$ of $\mathscr{A}$ and

$$(\mathscr{A}^{-1})^* = (\mathscr{A}^*)^{-1}.$$

6.15 Symmetric Operators

A linear operator $\mathscr{A}$ is said to be *symmetric* (or *self-adjoint*) if $\mathscr{A}$ is identical to its adjoint operator $\mathscr{A}^*$ so that $\mathscr{A}^* = \mathscr{A}$.

Notice that by virtue of Property (f) the matrix of a symmetric operator relative to an orthonormal basis is also symmetric and remains unchanged when being transposed. Therefore a symmetric operator is adjoint to itself and may also be called the self-adjoint operator.

Example. Let $\mathscr{P}$ be a linear operator that defines an orthogonal projection of a three-dimensional Euclidean space with the induced Cartesian coordinate system *Oxyz* onto the *xy*-plane (Fig. 6.18).

Relative to the orthonormal basis **i**, **j**, **k** the matrix of $\mathscr{P}$ takes the symmetric form

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

since $\mathscr{P}\mathbf{i} = \mathbf{i}$, $\mathscr{P}\mathbf{j} = \mathbf{j}$ and $\mathscr{P}\mathbf{k} = \boldsymbol{\theta}$. This means that the operator $\mathscr{P}$ is symmetric.

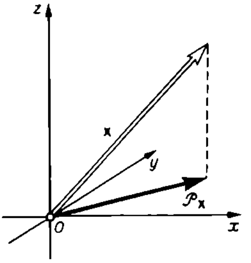


Fig. 6.18

Properties of symmetric operators. It is worth mentioning some remarkable properties of symmetric operators. The first two properties given below are the direct consequences of the definition of a symmetric operator.

(a) For a linear operator $\mathscr{A}$: $\mathbf{V} \to \mathbf{V}$ to be symmetric it is necessary and sufficient that for any vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbf{V}$ there holds

$$(\mathscr{A}\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathscr{A}\mathbf{y}). \tag{6.32}$$

(b) For a linear operator to be symmetric it is necessary and sufficient that relative to an orthonormal basis its matrix become symmetric.

(c) The characteristic polynomial of a symmetric operator (and the associated symmetric matrix) has only real roots.

Recall that any real root λ of the characteristic polynomial is the eigenvalue of the corresponding linear operator $\mathscr{A}$, i.e., there exists a nonzero vector $\mathbf{x}$ (the eigenvector of $\mathscr{A}$) such that $\mathscr{A}\mathbf{x} = \lambda\mathbf{x}$.

(d) The eigenvectors of a symmetric operator corresponding to distinct eigenvalues are orthogonal.

◄ Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be the eigenvectors of $\mathscr{A}$ so that $\mathscr{A}\mathbf{x}_1 = \lambda_1\mathbf{x}_1$ and $\mathscr{A}\mathbf{x}_2 = \lambda_2\mathbf{x}_2$, and let $\lambda_1 \neq \lambda_2$.

Since $\mathscr{A}$ is a symmetric operator we have

$$(\mathscr{A}\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1, \mathscr{A}\mathbf{x}_2).$$

On the other hand,

$$(\mathscr{A}\mathbf{x}_1, \mathbf{x}_2) = (\lambda_1\mathbf{x}_1, \mathbf{x}_2) = \lambda_1(\mathbf{x}_1, \mathbf{x}_2)$$

and

$$(\mathbf{x}_1, \mathscr{A}\mathbf{x}_2) = (\mathbf{x}_1, \lambda_2\mathbf{x}_2) = \lambda_2(\mathbf{x}_1, \mathbf{x}_2).$$

Whence

$$\lambda_1(\mathbf{x}_1, \mathbf{x}_2) = \lambda_2(\mathbf{x}_1, \mathbf{x}_2)$$

and $(\lambda_1 - \lambda_2)(\mathbf{x}_1, \mathbf{x}_2) = 0$.

Since $\lambda_1 - \lambda_2 \neq 0$ we arrive at

$(\mathbf{x}_1, \mathbf{x}_2) = 0$. ►

(e) Let $\mathscr{A}$: $\mathbf{V} \to \mathbf{V}$ be a symmetric operator. Then in $\mathbf{V}$ there exists an orthonormal basis $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ comprising the eigenvectors of $\mathscr{A}$ so that

$$\mathscr{A}\mathbf{e}_i = \lambda_i\mathbf{e}_i \quad (i = 1, 2, \ldots, n),$$
$$(\mathbf{e}_i, \mathbf{e}_j) = \delta_{ij} \quad (i, j = 1, 2, \ldots, n).$$

Turning back to the previous example we easily see that the triple (**i**, **j**, **k**) is the desired orthonormal basis in $\mathbf{V}$ since the vectors **i** and **j** are the eigenvectors of $\mathscr{P}$ corresponding to the eigenvalue 1 (of multiplicity 2) and **k** is the eigenvector corresponding to the eigenvalue 0.

(f) If a nonsingular operator $\mathscr{A}$: $\mathbf{V} \to \mathbf{V}$ is symmetric so is its inverse $\mathscr{A}^{-1}$: $\mathbf{V} \to \mathbf{V}$.

Remark. All the eigenvalues of a nonsingular operator are distinct from zero. Indeed, if $\lambda \neq 0$ is the eigenvalue of a nonsingular operator $\mathscr{A}$, then $1/\lambda$ is the eigenvalue of the inverse operator $\mathscr{A}^{-1}$.

We shall say that a symmetric operator $\mathscr{A}$ is *positive* if given any nonzero vector $\mathbf{x}$ in $\mathbf{V}$ there holds $(\mathscr{A}\mathbf{x}, \mathbf{x}) > 0$.

Properties of positive operators. (a) A symmetric operator $\mathscr{A}$: $\mathbf{V} \to \mathbf{V}$ is positive if and only if all the eigenvalues of $\mathscr{A}$ are positive.

(b) A positive operator is nonsingular.

(c) If an operator $\mathscr{A}$ is positive so is its inverse.

6.16 Quadratic Forms

Let $\mathbf{A} = (\alpha_{ij})$ be a symmetric matrix of order n so that $\alpha_{ji} = \alpha_{ij}$. Then the expression

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ij}\xi^i\xi^j \tag{6.33}$$

is said to be the *quadratic form* in the variables $\xi^1, \xi^2, \ldots, \xi^n$. The matrix $\mathbf{A}$ is called the *associated matrix* of the quadratic form.

The quadratic trinomial $ax^2 + 2bxy + cy^2$, where a, b and c are real numbers, serves as an example of the quadratic form in two variables x and y, the associated matrix being $\begin{pmatrix} a & b \\ b & c \end{pmatrix}$.

The n-tuple of numbers $\xi^1, \xi^2, \ldots, \xi^n$ may be regarded as the coordinates of a vector $\mathbf{x}$ in an n-dimensional real space $\mathbf{V}$ relative to a given basis so that

$$\mathbf{x} = \xi^1\mathbf{e}_1 + \xi^2\mathbf{e}_2 + \ldots + \xi^n\mathbf{e}_n,$$

where $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ is the orthonormal basis of $\mathbf{V}$. Then (6.33) becomes a numeric function of a vector-valued argument $\mathbf{x}$ defined over the space $\mathbf{V}$. This function is customarily written as

$$\mathscr{A}(\mathbf{x}, \mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ij}\xi^i\xi^j. \tag{6.34}$$

We shall also say that the function $\mathscr{A}(\mathbf{x}, \mathbf{x})$ is defined in an n-dimensional Euclidean space $\mathbf{V}$.

We may also associate with any quadratic form $\mathscr{A}(\mathbf{x}, \mathbf{x})$ the *bilinear form*

$$\mathscr{A}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ij}\xi^i\eta^j, \tag{6.35}$$

where $\eta^1, \eta^2, \ldots, \eta^n$ are the coordinates of the vector $\mathbf{y}$ relative to the orthonormal basis $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ so that

$$\mathbf{y} = \eta^1\mathbf{e}_1 + \eta^2\mathbf{e}_2 + \ldots + \eta^n\mathbf{e}_n.$$

The form (6.35) is called bilinear since it is linear in both the argument $\mathbf{x}$ and the argument $\mathbf{y}$ so that

$$\mathscr{A}(\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2, \mathbf{y}) = \alpha_1\mathscr{A}(\mathbf{x}_1, \mathbf{y}) + \alpha_2\mathscr{A}(\mathbf{x}_2, \mathbf{y})$$

and

$$\mathscr{A}(\mathbf{x}, \beta_1\mathbf{y}_1 + \beta_2\mathbf{y}_2) = \beta_1\mathscr{A}(\mathbf{x}, \mathbf{y}_1) + \beta_2\mathscr{A}(\mathbf{x}, \mathbf{y}_2),$$

where $\alpha_1, \alpha_2, \beta_1$ and β_2 are arbitrary numbers.

The bilinear form (6.35) is symmetric since its value is independent of the order in which $\mathbf{x}$ and $\mathbf{y}$ occur in (6.35), i.e.,

$$\mathscr{A}(\mathbf{y}, \mathbf{x}) = \mathscr{A}(\mathbf{x}, \mathbf{y}).$$

Computing the value of $\mathscr{A}(\mathbf{x}, \mathbf{y})$ for the base vectors, i.e., for $\mathbf{x} = \mathbf{e}_k$, $\mathbf{y} = \mathbf{e}_m$, we obtain

$$\mathscr{A}(\mathbf{e}_k, \mathbf{e}_m) = \alpha_{km}. \tag{6.36}$$

Whence it follows that the elements of the associated matrix $\mathbf{A}$ of the quadratic form (6.34) are the values of the bilinear form computed for the vectors of the basis $\mathbf{e}$.

The scalar product of two vectors in an n-dimensional coordinate space $\mathbb{R}^n$

$$(\xi, \eta) = \xi^1\eta^1 + \xi^2\eta^2 + \ldots + \xi^n\eta^n,$$

where $\xi = (\xi^1, \xi^2, \ldots, \xi^n) \in \mathbb{R}^n$ and $\eta = (\eta^1, \eta^2, \ldots, \eta^n) \in \mathbb{R}^n$, is the bilinear form.

The associated quadratic form

$$|\xi|^2 = (\xi, \xi) = (\xi^1)^2 + (\xi^2)^2 + \ldots + (\xi^n)^2$$

defines the square of the length of the vector ξ.

The coordinates of a vector $\mathbf{x}$ relative to a different basis are different and so is the matrix of the quadratic form.

In a variety of applications we need to simplify a quadratic form by converting it to a *diagonal* or *normal* form.

A quadratic form is said to be of diagonal form if the coefficients in all $\xi^i \cdot \xi^j$ are equal to zero. In other words, a quadratic form $\mathscr{A}(\mathbf{x}, \mathbf{x})$ is of a diagonal form if $\alpha_{ij} = 0$ for all $i \neq j$ and

$$\mathscr{A}(\mathbf{x}, \mathbf{x}) = \alpha_{11}(\xi^1)^2 + \alpha_{22}(\xi^2)^2 + \ldots + \alpha_{nn}(\xi^n)^2.$$

The associated matrix is also diagonal, i.e.,

$$\begin{pmatrix} \alpha_{11} & & 0 \\ & \alpha_{22}\ldots & \\ 0 & & \alpha_{nn} \end{pmatrix}.$$

Theorem 6.16. *For any quadratic form defined over a Euclidean space there exists an orthonormal basis relative to which the associated matrix becomes diagonal.*

◄ To prove this theorem we shall use the arguments that follow from properties of symmetric operators.

We choose the orthonormal basis $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ and consider the linear operator $\mathscr{A}: \mathbf{V} \to \mathbf{V}$ such that, relative to $\mathbf{e}$, the matrix (α_j^i) of $\mathscr{A}$ is identical to the matrix (α_{ij}) of a given quadratic form, i.e., $\alpha_j^i = \alpha_{ij}$. Since (α_j^i) is symmetric so is the operator $\mathscr{A}$.

Let us compute $(\mathscr{A}\mathbf{x}, \mathbf{x})$. Since the basis $\mathbf{e}$ is orthonormal we have

$$(\mathscr{A}\mathbf{e}_i, \mathbf{e}_j) = \alpha_i^j = \alpha_{ij},$$

and

$$(\mathscr{A}\mathbf{x}, \mathbf{x}) = \left(\mathscr{A}\left(\sum_{i=1}^n \xi^i \mathbf{e}_i\right), \sum_{j=1}^n \xi^j \mathbf{e}_j\right)$$
$$= \sum_{i=1}^n \sum_{j=1}^n \xi^i \xi^j (\mathscr{A}\mathbf{e}_i, \mathbf{e}_j) = \sum_{i=1}^n \sum_{j=1}^n \alpha_{ij} \xi^i \xi^j = \mathscr{A}(\mathbf{x}, \mathbf{x}).$$

Whence we infer that the quadratic form $\mathscr{A}(\mathbf{x}, \mathbf{x})$ defined over a Euclidean linear space $\mathbf{V}$ and the symmetric operator $\mathscr{A}$ acting in $\mathbf{V}$ are related as

$$\mathscr{A}(\mathbf{x}, \mathbf{x}) = (\mathscr{A}\mathbf{x}, \mathbf{x}). \qquad (6.37)$$

Recall that for any symmetric operator and, in particular, for $\mathscr{A}$ in $\mathbf{V}$ there exists an orthonormal basis $\mathbf{f} = (\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n)$ comprising the eigenvectors of $\mathscr{A}$ so that

$$\mathscr{A}\mathbf{f}_k = \lambda_k\mathbf{f}_k \quad (k = 1, 2, \ldots, n); \quad (\mathbf{f}_k, \mathbf{f}_m) = \delta_{km}. \tag{6.38}$$

Notice that

$$(\mathscr{A}\mathbf{f}_k, \mathbf{f}_m) = (\lambda_k\mathbf{f}_k, \mathbf{f}_m) = \lambda_k\delta_{km} = \begin{cases} \lambda_k & (k = m), \\ 0 & (k \neq m). \end{cases}$$

Substituting the expansion of $\mathbf{x}$

$$\mathbf{x} = \sum_{k=1}^{n} \eta^k\mathbf{f}_k$$

relative to the basis $\mathbf{f} = (\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n)$, into $(\mathscr{A}\mathbf{x}, \mathbf{x})$, we have

$$(\mathscr{A}\mathbf{x}, \mathbf{x}) = \left(\mathscr{A}\left(\sum_{k=1}^{n} \eta^k\mathbf{f}_k\right), \sum_{m=1}^{n} \eta^m\mathbf{f}_m\right)$$
$$= \sum_{k=1}^{n}\sum_{m=1}^{n} \eta^k\eta^m(\mathscr{A}\mathbf{f}_k, \mathbf{f}_m) = \sum_{k=1}^{n} \lambda_k(\eta^k)^2.$$

Whence, (6.37) yields

$$\mathscr{A}(\mathbf{x}, \mathbf{x}) = \sum_{k=1}^{n} \lambda_k(\eta^k)^2. \tag{6.39}$$

Thus, the matrix $\mathbf{A}(\mathbf{f})$ of the original quadratic form becomes diagonal relative to the basis $\mathbf{f}$, so that

$$\mathbf{A}(\mathbf{f}) = \begin{pmatrix} \alpha_1 & & & 0 \\ & \alpha_2 & & \\ & & \ddots & \\ 0 & & & \alpha_n \end{pmatrix}. \blacktriangleright$$

We may convert a quadratic form to diagonal form without making explicit computations of the base vectors of $\mathbf{f}$. It suffices to compute the eigenvalues of the corresponding linear operator or, equivalently, the eigenvalues of the associated matrix $\mathbf{A} = (\alpha_{ij})$ counted with their multiplicities.

Example. Reduce the quadratic form

$$\mathscr{A}(\mathbf{x}, \mathbf{x}) = 2xy + 2yz + 2xz$$

to the diagonal form.

◀ The associated matrix is

$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

To find the eigenvalues we must solve

$$\begin{vmatrix} -\lambda & 1 & 1 \\ 1 & -\lambda & 1 \\ 1 & 1 & -\lambda \end{vmatrix} = -\lambda^3 + 3\lambda + 2 = 0$$

yielding $\lambda_1 = 2$ and $\lambda_{2,3} = -1$.

Thus we have

$$\mathscr{A}(\mathbf{x}, \mathbf{x}) = 2\tilde{x}^2 - \tilde{y}^2 - \tilde{z}^2.$$

It is much harder to compute the desired orthonormal basis. To this end we shall find the eigenvectors of the symmetric operator $\mathscr{A}$ that are identical to the eigenvectors of the matrix of the quadratic form $\mathscr{A}(\mathbf{x}, \mathbf{x})$.

Let $\lambda = 2$. Consider the homogeneous linear system specified by the coefficient matrix

$$\begin{pmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{pmatrix}.$$

We have

$$\begin{aligned} -2x + y + z &= 0, \\ x - 2y + z &= 0, \\ x + y - 2z &= 0. \end{aligned}$$

All solutions of the system are proportional to the vector $(1, 1, 1)'$. Hence, the unit vector is $\tilde{\mathbf{i}} = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)'$.

Let $\lambda = -1$. The homogeneous linear system defined by the coefficient matrix

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

has two linearly independent solutions and we have to choose them so that they become orthogonal.

The system reduces to the single equation $x + y + z = 0$. Then the desired solutions are $(1, -2, 1)'$ and $(1, 0, -1)'$ and the unit vectors are $\tilde{\mathbf{j}} = (1/\sqrt{6}, -2/\sqrt{6}, 1/\sqrt{6})'$ and $\tilde{\mathbf{k}} = (1/\sqrt{2}, 0, -1/\sqrt{2})'$.

It is easy to verify that both the vector $\tilde{\mathbf{j}}$ and the vector $\tilde{\mathbf{k}}$ are orthogonal to the vector $\tilde{\mathbf{i}}$. (Notice that this result also follows from Property (d) of a symmetric operator.)

Then the desired orthonormal basis comprises the vectors

$$\tilde{\mathbf{i}} = \frac{\mathbf{i} + \mathbf{j} + \mathbf{k}}{\sqrt{3}}, \quad \tilde{\mathbf{j}} = \frac{\mathbf{i} - 2\mathbf{j} + \mathbf{k}}{\sqrt{6}}, \quad \tilde{\mathbf{k}} = \frac{\mathbf{i} - \mathbf{k}}{\sqrt{2}}. \ ►$$

Remark. We may accept any n-dimensional Euclidean space as $\mathbf{V}$. However, of practical interest is a coordinate space $\mathbb{R}^n$ whose elements are all possible ordered n-tuples of real numbers $\xi = (\xi^1, \xi^2, \ldots, \xi^n)$. The basis of $\mathbb{R}^n$ comprises the vectors $(1, 0, \ldots, 0, 0)$, $(0, 1, \ldots, 0, 0)$, $\ldots$, $(0, 0, \ldots, 1, 0)$ relative to which the scalar product of two vectors $\xi = (\xi^1, \xi^2,$

$\ldots, \xi^n)$ and $\eta = (\eta^1, \eta^2, \ldots, \eta^n)$ is given by the formula

$$(\xi, \eta) = \xi^1\eta^1 + \xi^2\eta^2 + \ldots + \xi^n\eta^n.$$

We shall describe the procedure that enables us to choose the basis relative to which a given quadratic form specified over an n-dimensional coordinate space becomes diagonal.

◄ Let $\mathscr{A}(\mathbf{x}, \mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{ij}\xi^i\xi^j$ be a given quadratic form.

Step 1. Write down the associated matrix

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{n1} & \alpha_{n2} & \ldots & \alpha_{nn} \end{pmatrix}.$$

Step 2. Solve the polynomial equation

$$\begin{vmatrix} \alpha_{11} - t & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} - t & \ldots & \alpha_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{n1} & \alpha_{n2} & & \alpha_{nn} - t \end{vmatrix} = 0$$

yielding the eigenvalues of the associated matrix of $\mathscr{A}(\mathbf{x}, \mathbf{x})$.

Arrange the eigenvalues $\lambda_1 \leqslant \lambda_2 \ldots \leqslant \lambda_n$, their multiplicities counted. (Notice that all the eigenvalues are real since the matrix is symmetric.)

Step 3. Let λ be a root of multiplicity k. Then the homogeneous linear system specified by the coefficient matrix

$$\begin{pmatrix} \alpha_{11} - \lambda & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} - \lambda & \ldots & \alpha_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{n1} & \alpha_{n2} & \ldots & \alpha_{nn} - \lambda \end{pmatrix}$$

has exactly k linearly independent solutions that form the fundamental system of solutions. On normalizing the solutions we obtain k pairwise unit vectors.

Repeating this process for the other eigenvalues, we obtain exactly n pairwise orthogonal unit vectors that comprise the orthonormal basis $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n$ of $\mathbb{R}^n$. Notice that the vectors corresponding to the distinct eigenvalues are orthogonal by virtue of Property (d) of a symmetric operator.

Step 4. Write down $\mathscr{A}(\mathbf{x}, \mathbf{x})$, relative to the basis $\mathbf{f} = (\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n)$, in the diagonal form

$$\mathscr{A}(\mathbf{x}, \mathbf{x}) = \lambda_1(\eta^1)^2 + \lambda_2(\eta^2)^2 + \ldots + \lambda_n(\eta^n)^2,$$

where $\mathbf{x} = \eta^1\mathbf{f}_1 + \eta^2\mathbf{f}_2 + \ldots + \eta^n\mathbf{f}_n$. ►

Definition. The quadratic form

$$\mathscr{A}(\mathbf{x}, \mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{ij} \xi^i \xi^j \tag{6.40}$$

is called *positive-definite* if given any nonzero vector **x** or, equivalently, given any nonzero n-tuple $\xi^1, \xi^2, \ldots, \xi^n$, there holds

$$\mathscr{A}(\mathbf{x}, \mathbf{x}) > 0.$$

The scalar square of an arbitrary vector $\xi = (\xi^1, \xi^2, \ldots, \xi^n)$ of an n-dimensional coordinate space given by the formula

$$(\xi, \xi) = (\xi^1)^2 + (\xi^2)^2 + \ldots + (\xi^n)^2$$

is an example of a positive-definite quadratic form.

On reducing the positive-definite quadratic form $\mathscr{A}(\mathbf{x}, \mathbf{x})$ to the diagonal form, we have

$$\mathscr{A}(\mathbf{x}, \mathbf{x}) = \lambda_1(\eta^1)^2 + \lambda_2(\eta^2)^2 + \ldots + \lambda_n(\eta^n)^2,$$

where $\lambda_1 > 0, \lambda_2 > 0, \ldots, \lambda_n > 0$.

Criterion for a quadratic form to be positive-definite. The quadratic form $\mathscr{A}(\mathbf{x}, \mathbf{x})$ is positive-definite if and only if the leading minors of the associated matrix $\mathscr{A}$, i.e., the minors cut out of the left-hand upper corner of $\mathscr{A}$, are all positive, i.e.,

$$\alpha_{11} > 0, \quad \begin{vmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{12} & \alpha_{22} \end{vmatrix} > 0, \ \ldots \begin{vmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1k} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{1k} & \alpha_{2k} & \ldots & \alpha_{kk} \end{vmatrix} > 0,$$

$$\ldots \begin{vmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{12} & \alpha_{22} & \ldots & \alpha_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{1n} & \alpha_{2n} & \ldots & \alpha_{nn} \end{vmatrix} > 0.$$

Diagonalization of a quadratic form by completing the square. We shall explain the procedure which is useful to convert a quadratic form to a diagonal form and, in particular, to decide on definiteness of quadratic forms.

◄ Let

$$\mathscr{A}(\mathbf{x}, \mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{ij} \xi^i \xi^j$$

be a given quadratic form and let $\alpha_{11} \neq 0$.

By a simple algebra we can reduce the sum of all terms involving ξ^1 to the form

$$\alpha_{11}(\xi^1)^2 + 2\alpha_{12}\xi^1\xi^2 + \ldots + 2\alpha_{1n}\xi^1\xi^n$$
$$= \alpha_{11}\left((\xi^1)^2 + 2\,\frac{\alpha_{12}}{\alpha_{11}}\,\xi^1\xi^2 + \ldots + 2\,\frac{\alpha_{1n}}{\alpha_{11}}\,\xi^1\xi^n\right)$$
$$= \alpha_{11}\left(\xi^1 + \frac{\alpha_{12}}{\alpha_{11}}\,\xi^2 + \ldots + \frac{\alpha_{1n}}{\alpha_{11}}\,\xi^n\right)^2 - \sum_{i=2}^{n}\sum_{j=2}^{n}\frac{\alpha_{1i}\alpha_{1j}}{\alpha_{11}}\,\xi^i\xi^j.$$

Setting

$$\eta^1 = \xi^1 + \frac{\alpha_{12}}{\alpha_{11}}\,\xi^2 + \ldots + \frac{\alpha_{1n}}{\alpha_{11}}\,\xi^n,$$
$$\eta^k = \xi^k \quad (k = 2, 3, \ldots, n),$$

we obtain

$$\mathscr{A}(\mathbf{x}, \mathbf{x}) = \alpha_{11}(\eta^1)^2 + \sum_{i=2}^{n}\sum_{j=2}^{n}\alpha'_{ij}\eta^i\eta^j,$$

where

$$\alpha'_{ij} = \alpha_{ij} - \frac{\alpha_{1i}\alpha_{1j}}{\alpha_{11}}.$$

We look now at

$$\mathscr{A}_1(\mathbf{x}, \mathbf{x}) = \sum_{i=2}^{n}\sum_{j=2}^{n}\alpha'_{ij}\eta^i\eta^j.$$

It is easy to see that $\mathscr{A}_1(\mathbf{x}, \mathbf{x})$ is a quadratic form in $(n - 1)$ variables and can also be represented as the sum of the square of one variable and the quadratic form in the other $(n - 2)$ variables. Thus, repeating this process of "completing the square" we finally arrive at the desired diagonal form of $\mathscr{A}(\mathbf{x}, \mathbf{x})$.

If $\alpha_{11} = 0$ but α_{ii} $(2 \leqslant i \leqslant n)$ is distinct from zero, we start the process by completing the square of ξ^i.

Now we suppose that in $\mathscr{A}(\mathbf{x}, \mathbf{x})$ all the coefficients in squares of ξ^i $(i = 1, 2, \ldots, n)$ are equal to zero, i.e., $\alpha_{11} = \alpha_{12} = \ldots = \alpha_{ii} = \alpha_{nn} = 0$. Then by the substitution

$$\xi^1 = \eta^1 + \eta^2,$$
$$\xi^2 = \eta^1 - \eta^2,$$
$$\xi^k = \eta^k \quad (k = 3, 4, \ldots, n)$$

the quadratic form $\mathscr{A}(\mathbf{x}, \mathbf{x})$ is reduced so that we again have the general case. Indeed, by this substitution the term $2\alpha_{12}\xi^1\xi^2$ is reduced to $2\alpha_{12}(\eta^1)^2 - 2\alpha_{12}(\eta^2)^2$. ►

Example. By completing the square reduce the quadratic form

$$\mathscr{A}(\mathbf{x}, \mathbf{x}) = 2xy + 2yz + 2zx$$

to the diagonal form.

◄ By the substitution

$$x = \tilde{x} + \tilde{y}, \quad y = \tilde{x} - \tilde{y}, \quad z = \tilde{z}$$

$\mathscr{A}(\mathbf{x}, \mathbf{x})$ is reduced to the form

$$\begin{aligned}\mathscr{A}(\mathbf{x}, \mathbf{x}) &= 2\tilde{x}^2 - 2\tilde{y}^2 + 4\tilde{x}\tilde{z} = 2(\tilde{x}^2 + 2\tilde{x}\tilde{z}) - 2\tilde{y}^2 \\ &= 2(\tilde{x} + \tilde{z})^2 - 2\tilde{y}^2 - 2\tilde{z}^2.\end{aligned}$$

Set $\hat{x} = \tilde{x} + \tilde{z}$, $\hat{y} = \tilde{y}$, $\hat{z} = \tilde{z}$. Then

$$\mathscr{A}(\mathbf{x}, \mathbf{x}) = 2\hat{x}^2 - 2\hat{y}^2 - 2\hat{z}^2. \text{ ►}$$

Remark. The major shortcoming of the process of completing the square is that it involves the trasformations of coordinates which are not orthogonal, that is, the new coordinates taken in pairs are not orthogonal.

On comparing the diagonal forms of $2xy + 2yz + 2zx$ obtained by applying the procedure that involves identification of an orthonormal basis and the procedure of completing the square we easily see that in both cases the number of positive terms remains unchanged and so does the number of negative terms. This is an important property of quadratic forms called the *law of inertia* which states that for any quadratic form the number of positive terms remains the same in all its diagonal forms and so does the number of negative terms and the number of zero terms. Thus these numbers are independent of procedures applied to reduce a given quadratic form to a diagonal form.

6.17 Classification of Curves and Surfaces of the Second Order

We are now well equipped to turn back to the analysis of the general equations of curves and surfaces of the second order which we have encountered in Chap. 4.

Plane curves. The general equation of the second-order curve in the xy-plane is

$$ax^2 + 2bxy + cy^2 + 2dx + 2ey + f = 0,$$

where $a^2 + b^2 + c^2 > 0$.

The associated matrix of the quadratic form $ax^2 + 2bxy + cy^2$ is $\begin{pmatrix} a & b \\ b & c \end{pmatrix}$.

Computing the roots λ_1 and λ_2 of the characteristic polynomial and the corresponding orthonormal eigenvectors $\tilde{\mathbf{i}}$ and $\tilde{\mathbf{j}}$, we can use $\tilde{\mathbf{i}}$ and $\tilde{\mathbf{j}}$ as the unit vectors of the new coordinate axes, the $\tilde{x}$- and $\tilde{y}$-axes, respectively

(Fig. 6.19). Then the original equation becomes

$$\lambda_1\tilde{x}^2 + \lambda_2\tilde{y}^2 + 2\tilde{d}\tilde{x} + 2\tilde{e}\tilde{y} + f = 0.$$

Two cases have to be distinguished: (1) $\lambda_1 \cdot \lambda_2 \neq 0$ and (2) either λ_1 or λ_2 is equal to zero.

(1) By the translation

$$X = \tilde{x} + \frac{\tilde{d}}{\lambda_1}, \quad Y = \tilde{y} + \frac{\tilde{e}}{\lambda_2}$$

the equation is reduced to the form

$$\lambda_1 X^2 + \lambda_2 Y^2 + \tilde{f} = 0.$$

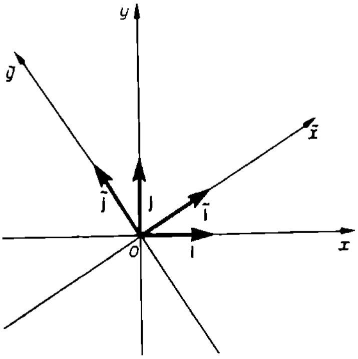


Fig. 6.19

In a way similar to that we have followed in Chap. 4 we consider all possible sequences of signs of λ_1, λ_2 and $\tilde{f}$ and finally arrive at the equations which specify an *ellipse*, a *hyperbola*, a *pair of intersecting lines*, a *point* and an *empty set* in the xy-plane.

(2) For definiteness we set $\lambda_1 = 0$ and $\lambda_2 \neq 0$. Then by the translation

$$X = \tilde{x} + \alpha, \quad Y = \tilde{y} + \frac{\tilde{e}}{\lambda_2}$$

the equation

$$\lambda_2\tilde{y}^2 + 2\tilde{d}\tilde{x} + 2\tilde{e}\tilde{y} + f = 0$$

is reduced to the equation

$$\lambda_2 Y^2 + 2\tilde{d}X + \tilde{f} = 0.$$

If $\tilde{d} \neq 0$ we put $\alpha = \dfrac{\tilde{f}}{2\tilde{d}}$, thus arriving at the equation of a *parabola*

$$\lambda_2 Y^2 + 2\tilde{d}X = 0.$$

If $\tilde{d} = 0$ we put $\alpha = 0$, thus obtaining the equation

$$\lambda_2 Y^2 + \tilde{f} = 0$$

which specifies either a *pair of parallel lines* or a *pair of coinciding lines* or an *empty set* corresponding to different signs of $\tilde{f}/\lambda_2$.

Remark. Computations of the roots of the characteristic polynomial and the corresponding orthonormal eigenvectors are used here instead of a suitable rotation of coordinate axes employed in Chap. 4 to eliminate the term $2bxy$ from the general equation.

Surfaces of the second order. The general equation is

$$\alpha_{11}x^2 + 2\alpha_{12}xy + 2\alpha_{13}xz + \alpha_{22}y^2 + 2\alpha_{23}yz + \alpha_{33}z^2 \\ + 2\alpha_{14}x + 2\alpha_{24}y + 2\alpha_{34}z + \alpha_{44} = 0,$$

where $\alpha_{11}^2 + \alpha_{12}^2 + \alpha_{13}^2 + \alpha_{22}^2 + \alpha_{23}^2 + \alpha_{33}^2 > 0$.

To simplify the quadratic form involved in the general equation we consider the associated matrix

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{12} & \alpha_{22} & \alpha_{23} \\ \alpha_{13} & \alpha_{23} & \alpha_{33} \end{pmatrix}$$

and compute the roots λ_1, λ_2 and λ_3 of the characteristic polynomial

$$\begin{vmatrix} \alpha_{11} - t & \alpha_{12} & \alpha_{13} \\ \alpha_{12} & \alpha_{22} - t & \alpha_{23} \\ \alpha_{13} & \alpha_{23} & \alpha_{33} - t \end{vmatrix}.$$

Let $\tilde{\mathbf{i}}$, $\tilde{\mathbf{j}}$ and $\tilde{\mathbf{k}}$ be the orthonormal eigenvectors of the associated matrix. We accept $\tilde{\mathbf{i}}$, $\tilde{\mathbf{j}}$, $\tilde{\mathbf{k}}$ as the unit vectors of the new coordinate axes, the $\tilde{x}$-, $\tilde{y}$- and $\tilde{z}$-axes, respectively, relative to which the general equation takes the form

$$\lambda_1\tilde{x}^2 + \lambda_2\tilde{y}^2 + \lambda_3\tilde{z}^2 + 2\tilde{\alpha}_{14}\tilde{x} + 2\tilde{\alpha}_{24}\tilde{y} + 2\tilde{\alpha}_{34}\tilde{z} + \alpha_{44} = 0.$$

Three cases have to be distinguished:

(1) All the three roots λ_1, λ_2, λ_3 are distinct from zero.

(2) Only one root is equal to zero (for definiteness we set $\lambda_3 = 0$).

(3) Two roots are equal to zero (for definiteness we set $\lambda_3 = \lambda_2 = 0$).

Remark. All the roots λ_1, λ_2, λ_3 are never equal to zero simultaneously.

We shall consider these three cases separately.

Case 1. By the translation

$$X = \tilde{x} + \frac{\tilde{\alpha}_{14}}{\lambda_1}, \quad Y = \tilde{y} + \frac{\tilde{\alpha}_{24}}{\lambda_2}, \quad Z = \tilde{z} + \frac{\tilde{\alpha}_{34}}{\lambda_3}$$

the general equation is reduced to the form

$$\lambda_1 X^2 + \lambda_2 Y^2 + \lambda_3 Z^2 + \tilde{\alpha}_{44} = 0.$$

(i) If $\tilde{\alpha}_{44} \neq 0$ and

(a) If λ_1, λ_2 and λ_3 are of the same sign which is opposite to that of $\tilde{\alpha}_{44}$, we have the equation of the *ellipsoid*

$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} + \frac{Z^2}{c^2} = 1,$$

where $a^2 = -\dfrac{\tilde{\alpha}_{44}}{\lambda_1}$, $b^2 = -\dfrac{\tilde{\alpha}_{44}}{\lambda_2}$, $c^2 = -\dfrac{\tilde{\alpha}_{44}}{\lambda_3}$.

(b) If the signs of λ_1 and λ_2 are opposite to that of $\tilde{\alpha}_{44}$ and λ_3 and $\tilde{\alpha}_{44}$ are of the same sign, we have the equation of the *hyperboloid of one sheet*

$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} - \frac{Z^2}{c^2} = 1,$$

where $a^2 = -\dfrac{\tilde{\alpha}_{44}}{\lambda_1}$, $b^2 = -\dfrac{\tilde{\alpha}_{44}}{\lambda_2}$, $c^2 = \dfrac{\tilde{\alpha}_{44}}{\lambda_3}$.

(c) If λ_1, λ_2 and $\tilde{\alpha}_{44}$ are of the same sign and the sign of λ_3 is opposite to that of $\tilde{\alpha}_{44}$, we have the equation of the *hyperboloid of two sheets*

$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} - \frac{Z^2}{c^2} = -1,$$

where $a^2 = \dfrac{\tilde{\alpha}_{44}}{\lambda_1}$, $b^2 = \dfrac{\tilde{\alpha}_{44}}{\lambda_2}$, $c^2 = -\dfrac{\tilde{\alpha}_{44}}{\lambda_3}$.

(ii) If $\tilde{\alpha}_{44} = 0$ and

(a) If λ_1, λ_2 and λ_3 are of the same sign, the equation defines the *point* (0, 0, 0) in space.

(b) If any two roots are of the same sign which is opposite to the sign of the third root, we have the equation of the *cone of the second order*

$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} - \frac{Z^2}{c^2} = 0.$$

Case 2. By the translation

$$X = \tilde{x} + \frac{\tilde{\alpha}_{14}}{\lambda_1}, \quad Y = \tilde{y} + \frac{\tilde{\alpha}_{24}}{\lambda_2}, \quad Z = \tilde{z}$$

the general equation is reduced to the form

$$\lambda_1 X^2 + \lambda_2 Y^2 + 2\tilde{\alpha}_{34} Z + \tilde{\alpha}_{44} = 0. \tag{*}$$

(i) If $\tilde{\alpha}_{34} \neq 0$, then by the translation

$$x = X,\ y = Y,\ z = Z + \frac{\tilde{\alpha}_{44}}{2\tilde{\alpha}_{34}}$$

the equation (*) is reduced to the form

$$\lambda_1 x^2 + \lambda_2 y^2 + 2\tilde{\alpha}_{34} z = 0,$$

in which case

(a) If λ_1 and λ_2 are of the same sign, we have the equation of the *elliptic paraboloid*

$$2z = \frac{x^2}{p} + \frac{y^2}{q},$$

where $p = -\frac{\tilde{\alpha}_{34}}{\lambda_1} > 0$ and $q = -\frac{\tilde{\alpha}_{34}}{\lambda_2} > 0$.

Notice that we have assumed that the sign of $\tilde{\alpha}_{34}$ is opposite to that of λ_1 and λ_2. We can always make this assumption by reflecting the z-axis, if necessary.

(b) If λ_1 and λ_2 are of the opposite signs, we have the equation of the *hyperbolic paraboloid*

$$2z = \frac{x^2}{p} - \frac{y^2}{q},$$

where $p = -\frac{\tilde{\alpha}_{34}}{\lambda_1} > 0$ and $q = +\frac{\tilde{\alpha}_{34}}{\lambda_2} > 0$.

(ii) If $\tilde{\alpha}_{34} = 0$ then the equation (*) becomes

$$\lambda_1 X^2 + \lambda_2 Y^2 + \tilde{\alpha}_{44} = 0,$$

which defines a family of *cylindrical surfaces* whose directing lines lie in the XY-plane and are given by the equation

$$\lambda_1 X^2 + \lambda_2 Y^2 + \tilde{\alpha}_{44} = 0.$$

The following table shows the classification of the corresponding cylindrical surfaces.

Case 3. By the translation

$$X = \tilde{x} + \frac{\tilde{\alpha}_{14}}{\lambda_1},\ Y = \tilde{y},\ Z = \tilde{z}$$

the general equation is reduced to the form

$$\lambda_1 X^2 + 2\tilde{\alpha}_{24} Y + 2\tilde{\alpha}_{34} Z + \tilde{\alpha}_{44} = 0.$$

(i) If $\tilde{\alpha}_{24} \neq 0$ and $\tilde{\alpha}_{34} \neq 0$ then we can always reduce this equation to the form with $\alpha_{24} \neq 0$ and $\alpha_{34} = 0$. It suffices to transform the coordinates

Table 6.1

$\lambda_1 \cdot \lambda_2 > 0$	$\lambda_1 \cdot \tilde{\alpha}_{44} > 0$	$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$	Elliptic cylinder
$\lambda_1 \cdot \lambda_2 > 0$	$\lambda_1 \cdot \tilde{\alpha}_{44} < 0$	$\frac{x^2}{a^2} + \frac{y^2}{b^2} = -1$	Empty set
$\lambda_1 \cdot \lambda_2 > 0$	$\tilde{\alpha}_{44} = 0$	$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 0$	z-axis
$\lambda_1 \cdot \lambda_2 < 0$	$\tilde{\alpha}_{44} \neq 0$	$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$	Hyperbolic cylinder
$\lambda_1 \cdot \lambda_2 < 0$	$\tilde{\alpha}_{44} = 0$	$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 0$	Pair of intersecting lines

as

$$\hat{x} = X, \quad \hat{y} = \frac{\tilde{\alpha}_{24} Y + \tilde{\alpha}_{34} Z}{\sqrt{\tilde{\alpha}_{24}^2 + \tilde{\alpha}_{34}^2}}, \quad \hat{z} = \frac{-\tilde{\alpha}_{34} Y + \tilde{\alpha}_{24} Z}{\sqrt{\tilde{\alpha}_{24}^2 + \tilde{\alpha}_{34}^2}}.$$

The equation becomes

$$\lambda_1 \hat{x}^2 + 2\beta \hat{y} + \tilde{\alpha}_{44} = 0, \tag{$*$}$$

where $\beta = \sqrt{\tilde{\alpha}_{24}^2 + \tilde{\alpha}_{34}^2}$.

Notice that we have to choose the transformation of coordinates such that the new system of coordinates will be the Cartesian one.

By the translation

$$x = \hat{x}, \quad y = \hat{y} + \frac{\tilde{\alpha}_{44}}{2\beta}, \quad z = \hat{z}$$

the equation $(*)$ is transformed to the equation of the *parabolic cylinder*

$$x^2 = 2py.$$

(ii) If $\alpha_{24} = \alpha_{34} = 0$ we have the equation

$$\lambda_1 X^2 + \tilde{\alpha}_{44} = 0,$$

which defines a *pair of parallel planes* when $\lambda_1 \cdot \tilde{\alpha}_{44} < 0$, a *pair of coinciding planes* when $\tilde{\alpha}_{44} = 0$ or an *empty set* when $\lambda_1 \cdot \tilde{\alpha}_{44} > 0$.

Exercises

1. Define the linear span generated by the polynomials $1 + t^2$, $t + t^2$, $1 + t + t^2$.

2. Determine whether the vectors $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ are linearly dependent or not: (a) $\mathbf{x}_1 = (1, 2, 3)$, $\mathbf{x}_2 = (4, 5, 6)$, $\mathbf{x}_3 = (7, 8, 9)$, (b) $\mathbf{x}_1 = (1, 4, 7, 10)$, $\mathbf{x}_2 = (2, 5, 8, 11)$, $\mathbf{x}_3 = (3, 6, 9, 12)$.

3. Show that the vectors $\mathbf{x}_1 = (1, 1, 1)$, $\mathbf{x}_2 = (1, 1, 0)$, $\mathbf{x}_3 = (0, 1, -1)$ form the basis of the linear space $\mathbb{R}^3$.

4. Complement the collection of two vectors $(1, 1, 0, 0)$ and $(0, 0, 1, 1)$ to get the basis of the linear space $\mathbb{R}^4$.

5. Verify that the vectors $(2, 2, -1)$, $(2, -1, 2)$, $(-1, 2, 2)$ form a basis of the linear space $\mathbb{R}^3$ and find the coordinates of the vector $\mathbf{x} = (3, 3, 3)$ relative to the basis of $\mathbb{R}^3$.

6. Define the dimension and a basis of the linear span generated by the vectors $\mathbf{x}_1 = (1, 2, 2, -1)$, $\mathbf{x}_2 = (2, 3, 2, 5)$, $\mathbf{x}_3 = (-1, 4, 3, -1)$, $\mathbf{x}_4 = (2, 9, 3, 5)$ of the linear space $\mathbb{R}^4$.

7. Compute the angle between the vectors $(2, -1, 3, -2)$ and $(3, 1, 5, 1)$ in the Euclidean space $\mathbb{R}^4$.

8. Apply the procedure of orthogonalization to the vectors $(1, -2, 2)$, $(-1, 0, -1)$ and $(5, -3, -7)$ in $\mathbb{R}^3$.

9. Let $\mathbf{L}$ be a subspace generated by the vectors $(1, 3, 3, 5)$, $(1, 3, -5, -3)$, $(1, -5, 3, -3)$. Find the $\mathbf{L}$-component of the vector $\mathbf{x}$ and the orthogonal complement of $\mathbf{x}$ with respect to $\mathbf{L}$ if $\mathbf{x} = (2, -5, 3, 4)$.

10. Let $\mathcal{A}$ be an operator defined over the 3-dimensional Euclidean space, such that $\mathcal{A}\mathbf{x} = (\mathbf{x}, \mathbf{a})\mathbf{a}$, where $\mathbf{a}$ is a given vector. Prove that $\mathcal{A}$ is a linear operator.

11. Let $\mathcal{A}$ be a linear operator that maps an arbitrary vector $\mathbf{x} = (x^1, x^2, x^3)$ as $\mathcal{A}\mathbf{x} = (2x^1 - x^2 - x^3, x^1 - 2x^2 + x^3, x^1 + x^2 - 2x^3)$. Find the image, kernel, rank and nullity of $\mathcal{A}$

12. Find the matrix of the differential operator defined over the 2-dimensional linear space generated by the base functions

$$\varphi(t) = e^t \cos t \quad \text{and} \quad \psi(t) = e^t \sin t.$$

13. Let

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

be a matrix of an operator $\mathcal{A}$ relative to the basis $\mathbf{1}$, $\mathbf{t}$, $\mathbf{t}^2$ of the space $\mathbf{M}_2$. Find the matrix of $\mathcal{A}$ relative to the basis formed by the polynomials $3t^2 + 2t$, $5t^2 + 3t + 1$, $7t^2 + 5t + 3$.

14. Compute the eigenvectors and the eigenvalues of the operators defined by the matrices

(a) $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$, (b) $\begin{pmatrix} 4 & -1 & -2 \\ 2 & 1 & -2 \\ 1 & -1 & 1 \end{pmatrix}$.

15. Let an operator define the rotation of a plane through the angle $\frac{\pi}{2}$. Find the operator adjoint to the given one.

16. Convert the quadratic form $2x^2 + 5y^2 + 2z^2 - 4xy - 2xz + 4yz$ to the diagonal form.

17. Specify what surfaces are given by the equations

(a) $7x^2 + 6y^2 + 5z^2 - 4xy - 4yz - 6x - 24y - 18z + 30 = 0$;

(b) $x^2 + 5y^2 + z^2 + 2xy + 6xz + 2yz - 2x + 6y + 2z = 0$;

(c) $5x^2 - y^2 + z^2 + 4xy + 6xz + 2x + 4y + 6z - 8 = 0$.

Answers

1. The collection of polynomials of order not exceeding 2. **2.** (a) yes; (b) yes. **4.** For example, (0, 1, 0, 0), (0, 0, 1, 0). **5.** (1, 1, 1). **6.** 4; $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$. **7.** $\pi/4$. **8.** $\left(\frac{1}{3}, -\frac{2}{3}, \frac{2}{3}\right)$, $\left(-\frac{2}{3}, -\frac{2}{3}, -\frac{1}{3}\right)$, $\left(\frac{2}{3}, -\frac{1}{3}, -\frac{2}{3}\right)$. **9.** $\mathbf{y} = (0, -3, 5, 2)$, $\mathbf{z} = (2, -2, -2, 2)$. **11.** The basis of the image is $\mathbf{y}_1 = (2, 1, 1)$, $\mathbf{y}_2 = (-1, 2, 1)$. The basis of the kernel is $\mathbf{z} = (1, 1, 1)$. The rank is 2. The nullity is 1. **12.** $\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$. **13.** $\begin{pmatrix} 1 & 0 & 0 \\ -15/4 & -4 & -5 \\ 9/4 & 3 & 4 \end{pmatrix}$.

14. (a) $\lambda_1 = \lambda_2 = 2$; (b) $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = 3$ and $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$; $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$, $\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$, $\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$.

15. The adjoint operator defines the rotation through the angle $-\frac{\pi}{2}$. **16.** $X^2 + 7Y^2 + Z^2$; $x = \frac{1}{\sqrt{2}} X + \frac{1}{\sqrt{6}} Y + \frac{1}{\sqrt{3}} Z$; $y = -\frac{2}{\sqrt{6}} Y + \frac{1}{\sqrt{3}} Z$; $z = \frac{1}{\sqrt{2}} X - \frac{1}{\sqrt{6}} Y - \frac{1}{\sqrt{3}} Z$.

17. (a) ellipsoid $\frac{x^2}{2} + \frac{y^2}{1} + \frac{z^2}{2/3} = 1$; (b) hyperboloid of one sheet $\frac{X^2}{1/3} + \frac{Y^2}{1/6} - \frac{Z^2}{1/12} = 1$; (c) hyperbolic paraboloid $\frac{X^2}{4/7\sqrt{14}} - \frac{Y^2}{2/\sqrt{14}} = 2Z$.

Chapter 7

An Introduction to Analysis

7.1 Basic Concepts

Sets. *Set* is a basic undefined concept in mathematics. We shall be content with the understanding that a set is a group or a collection of well-defined distinguishable objects which are thought of as a whole. It may be a set of letters printed on this page, a set of grains of sand on the seashore, a set of all roots of an equation or a set of all even numbers. Each object in a set is called an *element* or a *member* of the set. To signify that an element a is contained in a set A we write $a \in A$. The notation $a \notin A$ means that a does not belong to A.

Let A and B be two sets. If every element in A is also contained in B we say that A is a *subset* of B and write $A \subset B$. For instance, if Z is a set of all whole numbers and Z' is a set of all even numbers then $Z' \subset Z$. Notice that always $A \subset A$.

If $A \subset B$ and $B \subset A$, i.e., if every element in A is also contained in B and vice versa, we say that A and B are *equal* and write $A = B$. This means that a set is uniquely defined by its elements. So we may define a set by listing all the elements of the set enclosed in braces. The sets

$$A = \{a\},\ A = \{a, b\},\ A = \{a, b, c\}$$

consist of just one element a, two elements a and b, and three elements a, b and c, respectively. Sometimes it is impossible or impractical to list all elements of the set. In this case three dots are used to represent unlisted elements, e.g.,

$$A = \{a, b, c, \dots\}$$

is a set that consists of a, b, c and some other elements. To define the unlisted elements we shall use a written description that must fit all elements of the set and only elements of the set. For example, we shall write

the set of natural numbers $\{1, 2, 3, \dots\}$,
the set of squares of natural numbers $\{1, 4, 9, \dots\}$,
the set of primes $\{2, 3, 5, 7, \dots\}$.

If $A \subset B$ and $A \neq B$, A is called a *proper subset* of B.

Sometimes we do not know in advance whether a set contains at least a single element or not. So it is helpful to introduce the notion of an *empty set*, i.e., a set with no elements.[*] We shall denote the empty set by $\varnothing$. The empty set is a subset of any set or any set contains the empty set as its subset.

Operations on sets. Let A and B be two sets. The union of two sets A and B is the set $C = A \cup B$ of all elements contained either in A or in B, or in both. The intersection of two sets A and B is the set $C = A \cap B$ of all the elements contained both in A and in B. For example, let $A = \{1, 2, 3\}$ and $B = \{2, 3, 4, 5\}$. Then $A \cup B = \{1, 2, 3, 4, 5\}$ and $A \cap B = \{2, 3\}$.

If $A \cap B = \varnothing$, A and B are said to be *disjoint sets.*

We can similarly define the union and intersection of any number of sets.

Finite and infinite sets. A set is said to be *finite* if it has a *finite* number of elements. A set of all residents of a specific town and a set of people living on Earth are examples of finite sets.

A set is said to be *infinite* if it is not finite. The set $\mathbb{N} = \{1, 2, \ldots\}$ of all natural numbers is infinite.

Let A and B be two sets. We say that a *one-to-one correspondence* is set up between A and B if each element of A is associated with an element of B so that (i) distinct elements of A are associated with distinct elements of B and (ii) each element of B is put into correspondence with an element of A. It is easy to see that if a one-to-one correspondence is set up between the sets A and B these sets are called *equivalent.* We write $A \sim B$ to mean that A and B are equivalent sets.

An infinite set is said to be *countable* if it can be put into a one-to-one correspondence with the set $\mathbb{N}$ of natural numbers, i.e., if the set is equivalent to $\mathbb{N}$. *Any infinite set contains a countable subset.*

It can be known that the set of all rational numbers is countable while the set of all real numbers is uncountable.

Real numbers. The numbers 1, 2, 3, ... are called *natural* numbers. Every number which can be expressed as a fraction of the form $\pm \frac{m}{n}$, where m and n are natural numbers, and zero are *rational* numbers. Thus every positive integer and every negative integer are rational numbers. All rational numbers can be expressed as *repeating decimal fractions.* Unlike rational numbers, *irrational* numbers can be represented by *infinite nonrepeating decimal fractions.* The union of rational and irrational numbers forms a set of *real* numbers. It can be shown that the set of all rational numbers is countable while the set of all real numbers is uncountable. By convention the sets of natural, whole, rational and real numbers are denoted by $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$, respectively. We shall not give formal definitions of basic properties and operations on real numbers assuming that these are well familiar to the reader from the course of high-school mathematics.

---

[*] It has not yet been established whether a set of natural numbers n such that the equation $x^{n+2} + y^{n+2} = z^{n+2}$ has positive integral solutions is empty or not. (In other words, it has not yet been established if Fermat's last theorem is true or false.)

Absolute values of real numbers. Let a be a real number. The absolute value (or modulus) of a is equal to a if a is positive and is equal to $-a$ if a is negative. The absolute value of zero is zero. We denote the absolute value of a by $|a|$ and write

$$|a| = \begin{cases} a & \text{if } a \geqslant 0, \\ -a & \text{if } a < 0. \end{cases}$$

The inequality $|x| \leqslant \alpha$, where $\alpha > 0$, is equivalent to the relation

$$-\alpha \leqslant x \leqslant \alpha.$$

(Show that this relation is true.)

The basic properties of the absolute values are:

(1) $|a \cdot b| = |a| \cdot |b|$.

(2) $\left|\frac{a}{b}\right| = \frac{|a|}{|b|}$ $(b \neq 0)$.

◄ Relations (1) and (2) are direct consequences of laws of multiplication and division of real numbers and the definition of the absolute value of a real number. ►

(3) $|a + b| \leqslant |a| + |b|$.

◄ Indeed, it is easy to see that

$$-|a| \leqslant a \leqslant |a|,$$
$$-|b| \leqslant b \leqslant |b|.$$

Then adding the inequalities termwise, we obtain the inequality

$$-(|a| + |b|) \leqslant a + b \leqslant |a| + |b|,$$

which is equivalent to the desired relation $|a + b| \leqslant |a| + |b|$. ►

(4) $\big||a| - |b|\big| \leqslant |a - b|$.

◄ Indeed, the relation

$$|a| = |(a - b) + b| \leqslant |a - b| + |b|$$

implies that

$$|a| - |b| \leqslant |a - b|. \tag{$*$}$$

Similarly, the relation

$$|b| = |(b - a) + a| \leqslant |b - a| + |a| = |a - b| + |a|$$

yields

$$|a - b| \geqslant |b| - |a|$$

or

$$|a| - |b| \geqslant -|a - b|. \tag{$**$}$$

From (∗) and (∗∗) it follows that

$$-|a - b| \leqslant |a| - |b| \leqslant |a - b|.$$

Whence we obtain the desired inequality

$$\big||a| - |b|\big| \leqslant |a - b|. \quad \blacktriangleright$$

Absolute and relative errors. We shall introduce some notions that are widely used whenever we apply numerical methods to compute approximate solutions of mathematical problems.

Let a be a true value of some quantity and a^* be an approximation to a. We shall call a the exact number and a^* the approximate number. The simplest measure to estimate the precision of the approximate number a^* is the absolute error of a^*. We say that a positive number $\Delta(a^*)$ is the absolute error of a^* if

$$|a - a^*| \leqslant \Delta(a^*). \tag{$***$}$$

This definition of the absolute error is rather ambiguous. For example, if both a and a^* are known the absolute error of a^* is exactly equal to the absolute value of the difference between a and a^*. However, we may not know the value of a. In this case inequality (∗∗∗) means that the absolute value of the difference between a and a^* does not exceed $\Delta(a^*)$ and, consequently, any other positive number larger than $\Delta(a^*)$ may also be regarded as the absolute error of a^*.

The absolute error refers to the *precision* of approximation and tells us nothing about the *accuracy*. For example, if we have made two measurements of temperature with the same absolute error equal to 0.2 °C and found the readings 1000 °C and 10 °C, both measurements are to the same level of precision. However it is easy to see that the former is more accurate than the latter.

The accuracy of an approximation refers to the *relative* error. We say that a positive number $\delta(a^*)$ is the relative error of the approximation a^* to a if

$$\left|\frac{a - a^*}{a^*}\right| \leqslant \delta(a^*) \quad (a^* \neq 0).$$

The relative error is usually expressed as a percentage or a decimal fraction of a^*.

To signify that a^* is the approximation to a with the absolute error $\Delta(a^*)$ we write

$$a = a^* \pm \Delta(a^*),$$

where a^* and $\Delta(a^*)$ are expressed as decimal expansions with the same number of digits to the right of the decimal points. For example, the relation $a = 5.272 \pm 0.003$ means that $5.272 - 0.003 \leqslant a \leqslant 5.272 + 0.003$.

Similarly, if a^* is the approximation to a with the relative error $\delta(a^*)$ we write

$$a = a^*(1 \pm \delta(a^*)),$$

where $\delta(a^*)$ is expressed as a decimal fraction of a^*.

Significant digits. The first nonzero digit that occurs in the decimal expansion of the approximate number a^* on the left and all subsequent digits to the right of this digit are called the *significant digits*. For example, in $a_1^* = 0.\underline{2015}$, $a_2^* = \underline{23.653}$, $a_3^* = 0.00\underline{201500}$ the underlined digits are significant.

A significant digit is said to be *accurate* if the absolute error $\Delta(a^*)$ of the approximate number a^* does not exceed the number whose decimal expansion contains a unity in the decimal position corresponding to this significant digit. If in the decimal expansion of a^* the last accurate digit occupies the kth position, counting from the left, the approximate number a^* is said to be accurate to *k decimal positions*. For example, if $a^* = 0.3745$ and $\Delta(a^*) = 0.0001$ then all nonzero digits in a^* are significant. If $a = 23.6538$ and $a^* = 23.653$ then the approximate number is accurate to 3 decimal positions or to 5 digits, for the absolute error $\Delta(a^*) = 0.0008$ is smaller than the number obtained by substituting a unity for the last digit on the right and zeros for the other digits in a^*, i.e., $\Delta(a^*) = 0.0008 < 1 \times 10^{-3}$.

Sometimes it is helpful to apply a narrow definition of the accurate digit, namely, a significant digit is said to be accurate if the absolute error $\Delta(a^*)$ of a^* does not exceed the number whose decimal expansion contains half a unity in the decimal position corresponding to this significant digit. For example, if $a = 14.98$ then the approximate number $a^* = 15.00$ is accurate to 2 decimal positions or to 4 digits since $\Delta(a^*) = 0.02 < 0.5 \times 10^{-1}$.

The absolute and relative errors are usually represented by numbers with two or three significant digits.

Number line. Real numbers can be conveniently displayed as points on a line (Fig. 7.1).

Suppose that we have drawn a line and fixed the direction, the origin O and the unit distance e for this line. To display a real number a on the line we choose a point to the right of O such that the distance between this point and the origin O is equal to a if the number a is positive. When a is negative we choose a point to the left of the origin O at the distance equal to the absolute value $|a|$ of a. Clearly, the number $a = 0$ corresponds to the origin O. Therefore we have set up a one-to-one correspondence between the points of the line and the elements of the set of real numbers so that each real number is associated with one and only one point on the line and vice versa.

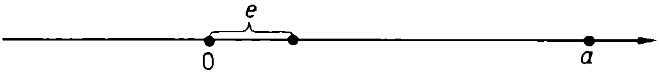


Fig. 7.1

The line whose points are put into a one-to-one correspondence with elements of the set of real numbers is called the *number line.*

We shall denote by x the point on the number line that corresponds to the real number x.

The number line is a mathematical model which is helpful to interpret relations between real numbers. The inequality $x_1 < x_2$ means that on the number line x_1 lies to the left of x_2. The inequality $x_1 < x_3 < x_2$ means that x_3 lies between x_1 and x_2.

Intervals. The important notions of intervals on the set of all real numbers are widely used in analysis.

A set of real numbers x is called

(a) the *closed interval* $[a, b]$ if the inequality $a \leqslant x \leqslant b$ holds for every x in the set;

(b) the *open interval* (a, b) if the inequality $a < x < b$ holds for every x in the set;

(c) the *half-open* or *half-closed* interval $(a, b]$ if the inequality $a < x \leqslant b$ holds for all x in the set. (Similarly, the set of real numbers x satisfying the inequality $a \leqslant x < b$ is also the half-open or half-closed interval denoted by $[a, b)$.)

We shall also consider *infinite* intervals by introducing points at infinity, $+\infty$ and $-\infty$ being the positive and negative infinities, respectively. For example

$(a, +\infty)$ is the set of real numbers such that $x > a$,
$[a, +\infty)$ is the set of real numbers such that $x \geqslant a$,
$(-\infty, b)$ is the set of real numbers such that $x < b$,
$(-\infty, b]$ is the set of real numbers such that $x \leqslant b$,
$(-\infty, +\infty)$ is the set $\mathbb{R}$ of all real numbers.

Neighbourhoods of points. Let x_0 be a point on the number line and $\delta > 0$ be a real number.

The interval which contains x_0 is called the *neighbourhood* of x_0. The interval $(x_0 - \delta, x_0 + \delta)$ which is symmetric relative to x_0 is called the *δ-neighbourhood* of x_0 (Fig. 7.2). Therefore the δ-neighbourhood of x_0 is the set of real numbers that satisfy the inequality $|x - x_0| < \delta$ or $x_0 - \delta < x < x_0 + \delta$. The interval $(x_0 - \delta, x_0 + \delta)$ which does not contain the point x_0 is called the *deleted δ-neighbourhood* of x_0.

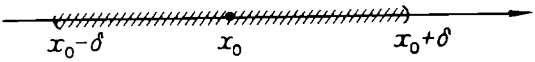


Fig. 7.2

Bounded and unbounded sets. Let E be a set of real numbers. Then the set E is called

(a) *bounded above* if there exists a number b such that $x \leqslant b$ for all $x \in E$;

(b) *bounded below* if there exists a number a such that $a \leqslant x$ for all $x \in E$;

(c) *bounded* if E is bounded above and bounded below, i.e., if there exist numbers a and b such that

$$a \leqslant x \leqslant b \quad \text{for all} \quad x \in E.$$

Therefore the set E is bounded if E is contained in the closed interval $[a, b]$.

For example, the set $E = (-\infty, 1]$ is bounded above and the set of all natural numbers is bounded below.

A set which is not bounded above (below) is called *unbounded above* (*below*). For example, the set of all natural numbers is unbounded above and bounded below. The set of all negative numbers is unbounded below and bounded above.

The sets of all integers, of all rational numbers and of all real numbers are unbounded.

Supremum of a set. Let E be a set bounded above, i.e., let there exist a number b such that $x \leqslant b$ for all $x \in E$. Then b is called an *upper bound* of E. Any number b' which is larger than b is also an upper bound of E.

Definition. The number M is called the *supremum* of E provided that

(i) for any $x \in E$ there holds $x \leqslant M$,

(ii) for any (whatever small) number $\varepsilon > 0$ there exists a number $x^* \in E$ such that $M - \varepsilon < x^* \leqslant M$.

In other words, the supremum of E is the *least upper bound* of E. We shall denote the supremum of E by $M = \sup E$ or $M = \sup\limits_{x \in E} \{x\}$.

If E is a set unbounded above we put the supremum of E equal to $+\infty$ and write $\sup E = +\infty$.

Infimum of a set. Let E be a set bounded below, i.e., let there exist a number a such that $a \leqslant x$ for all $x \in E$. Then a is called a *lower bound* of E. Clearly, any number smaller than a is also a lower bound of E.

Definition. The number m is called the *infimum* of E provided that

(i) for any $x \in E$ there holds $x \geqslant m$;

(ii) for any (whatever small) number $\varepsilon > 0$ there exists a number $x^* \in E$ such that $m \leqslant x^* < m + \varepsilon$.

Thus, the infimum of E is the *greatest lower bound* of E. We shall denote the infimum of E by $m = \inf E$ or $m = \inf\limits_{x \in E} \{x\}$.

If E is a set unbounded below we put the infimum of E equal to $-\infty$ and write $\inf E = -\infty$.

To illustrate the notions of supremum and infimum of a set we consider the following examples. If $E = [a, b]$ then $\inf E = a$ and $\sup E = b$. If $E = (a, b)$ then again $\inf E = a$ and $\sup E = b$. Notice that $\inf E$ and $\sup E$ are contained in E in the former example and do not belong to E in the latter one. For the set $E = \left\{1, \frac{1}{2}, \ldots, \frac{1}{n}, \ldots\right\}$ we have $\inf E = 0$ and $\sup E = 1$.

In conclusion we state the following theorem.

Theorem 7.1. *A non-empty set of real numbers which is bounded above has a supremum and a non-empty set of real numbers which is bounded below has an infimum.*

Logical symbols and connectives. To abbreviate the notation and simplify the definitions we shall use logical symbols and connectives.

The universal quantifier $\forall$ is read "For every", "For any", "For each", "For all".

The existential quantifier $\exists$ is read "There is", "There exist", "There exists at least one..." or something equivalent.

We say that a declarative sentence is a *statement* if it can be true or false but not both. For example, the sentences "Mathematics is a science", "The number 2 is smaller than the number 3" and "The number 6 is a prime" are statements whereas the sentences "Close the door" or "How old are you"? are not. We shall denote the statements by Greek letters α, β, γ, etc.

The implication $\alpha \Rightarrow \beta$ (read "if α then β" or "α implies β") is a false statement if and only if α, called the ***antecedent***, is true and β, called the ***consequent***, is false. We have to distinguish the implication from the ***cause-and-effect relation***. Unlike the latter the implication $\alpha \Rightarrow \beta$ is true whenever the antecedent α is false. In other words, a false statement implies any statement, e.g., "if $2 \times 2 = 5$ then the unidentified flying object has landed near your house".

The ***equivalence*** $\alpha \Leftrightarrow \beta$ (read "α if and only if β") means that the statements α and β are logically equivalent.

The ***conjunction*** $\alpha \wedge \beta$ (read "α and β") is a compound statement made up of the statements α and β connected by the conjunction ***and***. The conjunction $\alpha \wedge \beta$ is regarded as a true statement if and only if both α and β are true.

The ***disjunction*** $\alpha \vee \beta$ (read "α or β") is a compound statement made up of the statements α and β connected by the conjunction ***or***. The disjunction is thought of as true if and only if at least one of the statements is true.

Let α be a statement. The statement $\bar{\alpha}$ (read "not α’) is called the ***negation*** of α, $\bar{\alpha}$ being true if α is false and vice versa.

To negate a statement that involves the quantifiers we have to substitute every universal quantifier for the existentional quantifier and vice versa and replace the antecedent by its logical opposite so that if $\beta \Rightarrow \gamma$ then $\overline{\beta \Rightarrow \gamma} \Leftrightarrow \beta \wedge \bar{\gamma}$.

Necessary and sufficient conditions. Let the theorem "If the statement α is true so is the statement β" be true. The statements α and β that can be compound statements are called the ***hypothesis*** and ***conclusion***, respectively. The theorem can be symbolized as the implication $\alpha \Rightarrow \beta$ and can also be expressed as

α is a ***sufficient condition*** for β

or

β is a ***necessary condition*** for α.

Now we shall find out what we mean when speaking of the necessary and sufficient conditions.

Let β be a statement. We say that a statement α is a sufficient condition for β if α implies β and α is a necessary condition for β if α follows from β.

Let α and β be two statements given as follows

α: "The number x is equal to zero",

β: "The product xy is equal to zero".

Then $\alpha \Rightarrow \beta$.

◄ Indeed, for xy to be equal to zero it is sufficient that x be equal to zero. For x to be equal to zero it is necessary that xy be equal to zero.

But β is not a sufficient condition for α since x can be distinct from zero when xy is equal to zero. ►

If α and β are the statements each of which implies the other, i.e., $\alpha \Rightarrow \beta$ and $\beta \Rightarrow \alpha$, we say that each of α and β is the necessary and sufficient condition for the other and write

$$\alpha \Leftrightarrow \beta.$$

The following expressions all mean that α is the necessary and sufficient condition for β and vice versa:

(a) for α to be true it is necessary and sufficient that β hold;

(b) α holds if and only if β is satisfied;

(c) α is true if and only if β is true.

Mathematical induction. It is not a rare case when a statement which is true in some particular instances turns out to be false in general. For example, if we compute the values of $991n^2 + 1$ for the subsequent natural numbers 1, 2, 3, ..., 10^{10} we shall fail to get at least one value which is equal to the square of a natural number. Based upon this experience we might conjecture that the expression $991n^2 + 1$ will never produce squares of natural numbers when n is natural. However this conclusion would be false. The point is that the smallest n such that the value of $991n^2 + 1$ becomes equal to the square of a natural number is extremely large, viz.,

$$n = 12055735790331359447442538767.$$

Against this background it seems reasonable to draw our attention to the following problem. Let there be a statement which is true in some particular cases. How can we prove that it is true in general without having to verify it for each particular case that would be an impossible task?

An important tool which enables us to answer the imposed question is the *mathematical (complete) induction method* based on the principle of mathematical induction.

Principle of mathematical induction. Let α be a statement that is true for certain n. Then α is true for all natural n provided that

(a) α is true for $n = 1$;

(b) if α is true for $n = k$ then α is true for $n = k + 1$.

This principle lays down the basis of mathematical reasoning.

To illustrate how the principle of mathematical induction works we shall prove *Bernoulli's inequality:*

If $h > -1$, then

$$(1 + h)^n \geqslant 1 + nh \quad \forall n \in \mathbb{N}. \tag{*}$$

Clearly, inequality (*) is true for $n = 1$. Assume that (*) has been proved for $n = m > 1$, i.e.,

$$(1 + h)^m \geqslant 1 + mh.$$

Multiplying both sides of this inequality by $(1 + h) > 0$, we have

$$(1 + h)^{m+1} > (1 + mh)(1 + h) = 1 + (m + 1)h + mh^2.$$

Deleting the nonnegative number mh^2 from the right-hand side, we obtain

$$(1 + h)^{m+1} \geqslant 1 + (m + 1)h.$$

Thus $(*)$ is true for $n = m + 1$. Hence, Bernoulli's inequality $(*)$ is true for all $n \in \mathbb{N}$.

7.2 Sequences of Numbers

Notion and notation. Let every natural number n be associated with a real number a_n and let a rule that puts n into correspondence with a_n be known. Then we say that a *sequence of numbers*

$$a_1, a_2, \ldots, a_n, \ldots$$

is defined on the set of natural numbers.

The numbers $a_1, a_2, \ldots, a_n$ are called the *terms* of the sequence. We call a_n the *nth term of the sequence;* a_n is usually given by a law which enables us to compute any other term of the sequence.

To abbreviate the notation we shall denote the sequence $a_1, a_2, \ldots, a_n, \ldots$ by writing $\{a_n\}$*).

Examples of sequences are

$$\left\{\frac{1}{n}\right\} = 1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}, \ldots;$$

$$\{2^n\} = 2, 4, 8, \ldots, 2^n, \ldots;$$

$$\{1\} = 1, 1, 1, \ldots, 1, \ldots;$$

$$\left\{\frac{(-1)^{n+1}}{n^2}\right\} = 1, -\frac{1}{2^2}, \frac{1}{3^2}, \ldots, \frac{(-1)^{n+1}}{n^2}, \ldots;$$

$$\{\cos n\} = \cos 1, \cos 2, \cos 3, \ldots, \cos n, \ldots .$$

Limit of a sequence. A number A is said to be the *limit of a sequence* $\{a_n\}$ if, given any positive number ε, there exists a number N such that for all $n > N$

$$|a_n - A| < \varepsilon.$$

---

*) We have to distinguish the sequence $\{a_n\}$ from the set $\{a_n\}$. For example, $\{5\} = 5, 5, \ldots, 5, \ldots$ is a sequence whereas the set $\{5\}$ contains a single element, 5.

We shall write

$$A = \lim_{n \to \infty} a_n \quad \text{or} \quad a_n \xrightarrow[n \to \infty]{} A$$

to mean that A is the limit of the sequence $\{a_n\}$.

Using logical symbols we can write the definition of a limit as

$$(\lim_{n \to \infty} a_n = A) \Leftrightarrow \forall \varepsilon > 0 \quad \exists N \quad \forall n > N \Rightarrow |a_n - A| < \varepsilon. \tag{7.1}$$

The notion of a limit is easy to interpret by displaying the terms of the sequence $\{a_n\}$ on the number line (Fig. 7.3). The inequality $|a_n - A| < \varepsilon$ is equivalent to $A - \varepsilon < a_n < A + \varepsilon$ which implies that a_n lies in the ε-neighbourhood of A. Hence, A is the limit of $\{a_n\}$ if, given any ε-neighbourhood of A, there exists a number N such that all a_n with $n > N$ are contained in this ε-neighbourhood of A, i.e., in the open interval $(A - \varepsilon, A + \varepsilon)$. Thus, only the finite number of terms $a_1, a_2, \ldots, a_N$ can be outside the open interval $(A - \varepsilon, A + \varepsilon)$. Whence it follows that the sequence with all terms equal to A, called the *stationary* sequence, has the limit equal to A.

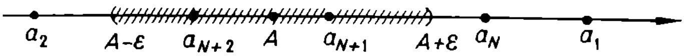


Fig. 7.3

A sequence is said to *converge* if it has a finite limit and to *diverge* otherwise.

Example. Consider the sequence $\left\{\frac{n+1}{n}\right\}$, $a_n = \frac{n+1}{n}$ being the nth term. Clearly, the larger n the closer to 1 is the fraction $\frac{n+1}{n} = 1 + \frac{1}{n}$. This prompts us to assume that

$$\lim_{n \to \infty} a_n = \lim_{n \to \infty} \frac{n+1}{n} = 1.$$

◄ To prove that our assumption is true we shall take an arbitrary $\varepsilon > 0$ and show that there exists a number N such that for all $n > N$

$$|a_n - 1| = \left| \frac{n+1}{n} - 1 \right| = \frac{1}{n} < \varepsilon. \tag{7.2}$$

From the inequality $\frac{1}{n} < \varepsilon$ we have $n > \frac{1}{\varepsilon}$. Then for any natural N exceeding $\frac{1}{\varepsilon}$ and all $n > N$ there holds $\frac{1}{n} < \varepsilon$, i.e., (7.2) is true. Hence, by virtue of the definiton of a limit we have $\lim_{n \to \infty} a_n = 1$. ►

In general the number N is not independent of ε, i.e., $N = N(\varepsilon)$. For instance, referring to the previous example, we observe that we can choose N equal to 10 or to any other number exceeding 10 when $\varepsilon = 0.1$. However, if $\varepsilon = 0.01$ N has to be larger than 100.

Remark. The number N is not uniquely defined by the value of ε in the sense that if inequality (7.1) holds for all $n > N_1$ it also holds for all $n > N_2$, where $N_2 > N_1$. Then to prove that $\lim\limits_{n\to\infty} a_n = A$ it suffices to choose any number N such that $|a_n - A| < \varepsilon$ for all $n > N$. Therefore we do not need to find the smallest number N satisfying (7.1).

How we shall prove two important theorems on limits of sequences.

Theorem 7.2 (Cauchy convergence criterion). *For the sequence* $a_1, a_2, \ldots, a_n \ldots$ *to converge it is necessary and sufficient that for any* $\varepsilon > 0$ *there exists* N *such that for all* $n > N$ *and all* $m > N$

$$|a_n - a_m| < \varepsilon.$$

A sequence $\{a_n\}$ satisfying Theorem 7.2 is said to be a *Cauchy sequence.*

Theorem 7.3 (on uniqueness of a limit). *A sequence* $\{a_n\}$ *can not have two distinct limits.*

◄ Let A be a limit of $\{a_n\}$ and let $B \neq A$. To prove that B is not a limit of $\{a_n\}$ we take ε so small that the ε-neighbourhood of A and the ε-neighbourhood of B do not intersect. For the purpose it suffices to set $\varepsilon = \dfrac{|E - A|}{3}$ (Fig. 7.4).

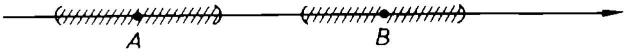


Fig. 7.4

Since $\lim\limits_{n\to\infty} a_n = A$, only a finite number of terms of $\{a_n\}$ can be outside the open interval $(A - \varepsilon, A + \varepsilon)$. Hence, the open interval $(B - \varepsilon, B + \varepsilon)$ contains at best a finite number of terms of $\{a_n\}$ and B is not a limit of $\{a_n\}$. ►

Bounded sequences. A sequence $\{a_n\}$ is called

(a) *bounded above* if there exists a number M such that $a_n \leqslant M$ for all n;

(b) *bounded below* if there exists a number m such that $a_n \geqslant m$ for all n;

(c) *bounded* if $\{a_n\}$ is bounded both above and below, i.e., there exist numbers m and M such that $m \leqslant a_n \leqslant M$ for all n.

Notice that all terms of a bounded sequence are contained in the closed interval $[m, M]$ of the number line.

It is easy to observe that the sequence $\ldots, -n, \ldots, -(n-1) \ldots, -3, -2, -1$ is bounded above, for all its terms are negative, whereas the sequence $1, 2, 3, \ldots, n, \ldots$ is bounded below since all its terms are not smaller than 1. The sequence $\{a_n\}$, where $a_n = \frac{n+1}{n} = 1 + \frac{1}{n}$ is the nth term, is bounded since $1 < a_n \leqslant 2$ for all n.

Sometimes it is helpful to use a slightly different definition of a bounded sequence, namely, a sequence $\{a_n\}$ is called *bounded* if there exists a number $K > 0$ such that

$$|a_n| \leqslant K \quad \forall n.$$

By using logical symbols we can write this definiton as

$$(\{a_n\} \text{ is bounded}) \Leftrightarrow \exists K > 0 \quad \forall n \quad |a_n| \leqslant K.$$

A definition of an unbounded sequence is easily obtained from the definition of a bounded sequence by interchanging the quantifiers and converting the inequality involved, i.e.,

$$(\{a_n\} \text{ is unbounded}) \Leftrightarrow \forall K > 0 \quad \exists n \quad |a_n| > K.$$

Example. Prove that the sequence $\{2^n\}$ is unbounded.

◄ Evidently, for any $K > 0$ there exists n such that $2^n > K$, i.e., $n > \log_2 K$. Hence, the sequence $\{2^n\}$ is unbounded. ►

A sequence is said to *diverge* to ∞, and we write $\lim\limits_{n \to \infty} a_n = \infty$, if, given any, whatever large, $M > 0$ there exists a number $N = N(M)$ such that

$$|a_n| > M \quad \forall n > N.$$

A sequence $\{a_n\}$ such that

$$\forall M > 0 \quad \exists N \quad \forall n > N \quad a_n > M \quad (a_n < -M)$$

is said to *diverge* to $+\infty$ $(-\infty)$. In this case we write $\lim\limits_{n \to \infty} a_n = +\infty$ $(\lim\limits_{n \to \infty} a_n = -\infty)$.

We shall say that a sequence which diverges to ∞, $+\infty$ or $-\infty$ is *infinitely large.* It has no finite limit satisfying the definition of a limit given above since the symbols ∞, $+\infty$ and $-\infty$ do not represent real numbers. In what follows we shall mean a finite limit as a real number. It is also worth mentioning here that there exist unbounded sequences which are not infinitely large. For example, the sequence $\left\{n \sin \frac{n\pi}{2}\right\}$ is unbounded but does not diverge to infinity.

Theorem 7.4. *Every convergent sequence is bounded, i.e., there exist numbers m and M such that for all n*

$$m \leqslant a_n \leqslant M.$$

◄ Let $\lim\limits_{n \to \infty} a_n = A$ and let $\varepsilon > 0$ be an arbitrary number. Then there exists

N such that the open interval $(A - \varepsilon, A + \varepsilon)$ contains all the terms a_n with $n > N$ and $a_1, a_2, \ldots, a_N$ are the only terms of the sequence which can lie outside this interval (Fig. 7.5). Thus only a finite number of terms can lie outside the interval and we can choose the smallest $\tilde{a}$ and the largest $\tilde{\tilde{a}}$ of them. Now let $m = \min\{\tilde{a}, A - \varepsilon\}$ be the smallest of the numbers $\tilde{a}$ and $A - \varepsilon$ and let $M = \max\{\tilde{\tilde{a}}, A + \varepsilon\}$ be the largest of the numbers $\tilde{\tilde{a}}$ and $A + \varepsilon$. Then the closed interval $[m, M]$ contains the terms $a_1, a_2, \ldots, a_N$ and the open interval $(A - \varepsilon, A + \varepsilon)$. Since all the terms a_n with $n \geqslant N + 1$ belong to $(A - \varepsilon, A + \varepsilon)$ the closed interval $[m, M]$ contains all the terms of the sequence $\{a_n\}$. Hence, $\{a_n\}$ is a bounded sequence. ►

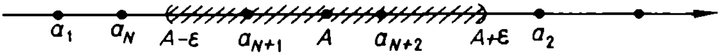


Fig. 7.5

Theorem 7.4 implies that for a sequence to converge it is necessary that this sequence be bounded. However, this is not sufficient for a sequence to converge, as is easily seen from the following example.

Example. The sequence

$$1, 0, 1, 0, 1, \ldots \qquad (*)$$

is bounded but diverges.

◄ To prove that the sequence diverges we suppose the converse, i.e., the sequence $(*)$ has a limit, say, equal to A. Then for any $\varepsilon > 0$ say for $\varepsilon = \frac{1}{4}$, there exists N such that

$$|a_n - A| < \frac{1}{4} \quad \forall n > N.$$

Since the terms of $(*)$ alternate between 0 and 1 there must hold

$$|0 - A| = |A| < \frac{1}{4} \quad \text{and} \quad |1 - A| < \frac{1}{4} \quad \forall n > N.$$

Whence, we have

$$1 = |(1 - A) + A| \leqslant |1 - A| + |A| < \frac{1}{4} + \frac{1}{4} = \frac{1}{2},$$

that is, $1 < \frac{1}{2}$, which is impossible. This means that our assumption on convergence of $(*)$ is false. Hence, the sequence $(*)$ does not have a limit, i.e., it diverges. ►

Operations on convergent sequences. We shall discuss a number of theorems which define arithmetic operations on convergent sequences and basic properties of limits of sequences.

Theorem 7.5. *Let* $\{a_n\}$ *and* $\{b_n\}$ *be sequences which converge to* A *and* B*, respectively. Then* $\{a_n + b_n\}$ *is a convergent sequence and*

$$\lim_{n\to\infty} (a_n + b_n) = A + B = \lim_{n\to\infty} a_n + \lim_{n\to\infty} b_n.$$

◄ Let $\varepsilon > 0$ be an arbitrary number. Since $\lim\limits_{n\to\infty} a_n = A$, there exists N_1 such that

$$|a_n - A| < \frac{\varepsilon}{2} \quad \forall n > N_1. \tag{**}$$

Similarly, since $\lim\limits_{n\to\infty} b_n = B$, there exists N_2 such that

$$|b_n - B| < \frac{\varepsilon}{2} \quad \forall n > N_2. \tag{***}$$

Now let N be the largest of N_1 and N_2. Then for any $n > N$ there hold both (**) and (***). Hence, we have

$$\begin{aligned} |(a_n + b_n) - (A + B)| &= |(a_n - A) + (b_n - B)| \\ &\leqslant |a_n - A| + |b_n - B| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

Thus

$$\forall \varepsilon > 0 \quad \exists N \quad \forall n > N \Rightarrow |(a_n + b_n) - (A + B)| < \varepsilon.$$

By virtue of the definition of a limit we infer that $A + B$ is the limit of the sequence $\{a_n + b_n\}$. ►

Theorem 7.5 can easily be generalized to any finite number of convergent sequences.

Analogously we can prove

Theorem 7.6. *If* $\{a_n\}$ *and* $\{b_n\}$ *are convergent sequences so is* $\{a_n - b_n\}$ *and*

$$\lim_{n\to\infty} (a_n - b_n) = \lim_{n\to\infty} a_n - \lim_{n\to\infty} b_n.$$

Also of importance are

Theorem 7.7. *If* $\{a_n\}$ *and* $\{b_n\}$ *are convergent sequences so is* $\{a_n \cdot b_n\}$ *and*

$$\lim_{n\to\infty} (a_n \cdot b_n) = \lim_{n\to\infty} a_n \cdot \lim_{n\to\infty} b_n.$$

Theorem 7.8. *If* $\{a_n\}$ *and* $\{b_n\}$ *are convergent sequences and if* $b_n \neq 0$ *for all* n *and* $\lim\limits_{n\to\infty} b_n \neq 0$ *then* $\left\{\dfrac{a_n}{b_n}\right\}$ *is a convergent sequence and*

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{\lim\limits_{n\to\infty} a_n}{\lim\limits_{n\to\infty} b_n}.$$

Monotone sequences. A sequence $\{a_n\}$ is called

(a) *nondecreasing* if $a_1 \leqslant a_2 \leqslant \ldots \leqslant a_n \leqslant a_{n+1} \leqslant \ldots$;

(b) *nonincreasing* if $a_1 \geqslant a_2 \geqslant \ldots \geqslant a_n \geqslant a_{n+1} \geqslant \ldots$;

(c) *monotone* if $\{a_n\}$ is either nondecreasing or nonincreasing.

A nondecreasing sequence $\{a_n\}$ is *bounded* if it is *bounded above,* i.e., if there exists a number M such that $a_n \leqslant M \ \forall n$, for all the terms of $\{a_n\}$ are contained in the closed interval $[a_1, M]$.

A nonincreasing sequence $\{a_n\}$ is *bounded* if it is *bounded below,* i.e., if there exists a number m such that $a_n \geqslant m \ \forall n$, for all the terms of $\{a_n\}$ are contained in the closed interval $[m, a_1]$.

Theorem 7.9. *Every monotone and bounded sequence has a limit.*

◄ Since a sequence $\{a_n\}$ is bounded the terms of $\{a_n\}$ form a set which has a supremum and an infimum. Let M be a supremum of this set and let us show that $\lim\limits_{n\to\infty} a_n = M$ provided that $\{a_n\}$ is a nondecreasing sequence.

By definition of a supremum for any $\varepsilon > 0$ there exists a_N such that $a_N > M - \varepsilon$ and $a_N < M$. Whence, we have $0 \leqslant M - a_N < \varepsilon$. Since $\{a_n\}$ is a nondecreasing sequence we have

$$0 \leqslant M - a_n \leqslant M - a_N \quad \forall n \geqslant N.$$

Then by virtue of $M - a_N < \varepsilon$, we get

$$0 \leqslant M - a_n < \varepsilon \ \forall n \geqslant N$$

or, equivalently,

$$|a_n - M| < \varepsilon \quad \forall n > N.$$

Whence it follows that M is a limit of $\{a_n\}$.

Analogously we prove that $\lim\limits_{n\to\infty} a_n = m$ provided that $\{a_n\}$ is a nonincreasing bounded sequence and m is an infimum of the set of the terms of $\{a_n\}$. ►

Remark. For a sequence to converge it is not necessary that this sequence be monotone. For example, the sequence $\left\{\frac{(-1)^{n+1}}{n}\right\}$ is not monotone but converges to 0, i.e., $\lim\limits_{n\to\infty} a_n = 0$.

From Theorem 7.9 follows the *nested closed interval theorem* sometimes called Cantor lemma.

Cantor lemma. Let there be given a sequence of the nested closed intervals

$$\sigma_n = [a_n, b_n] \quad (n = 1, 2, \ldots),$$

such that $\sigma_{n+1} \subset \sigma_n$ $(n = 1, 2, \ldots)$ and $d_n = b_n - a_n \to 0$ as $n \to \infty$. Then there exists only one point which belongs to all σ_n.

Cantor lemma refers to the remarkable property of a set of real numbers, i.e., the completeness of a number line, which implies that real numbers fill the number line without leaving "holes" in it.

Number e and natural logarithms. Let

$$a_n = \left(1 + \frac{1}{n}\right)^n \tag{$*$}$$

be the nth term of sequence $\{a_n\}$. Then substituting 1, 2 and 3 for n in ($*$), we get

$$a_1 = 2, \quad a_2 = \left(1 + \frac{1}{2}\right)^2 = 2 + \frac{1}{4}, \quad a_3 = \left(1 + \frac{1}{3}\right)^3 = 2 + \frac{1}{3} + \frac{1}{27}.$$

Thus $a_1 < a_2 < a_3$.

Using the binomial theorem*), we can easily show that the sequence $\left\{\left(1 + \frac{1}{n}\right)^n\right\}$ is a bounded and monotonically increasing sequence and

$$2 \leqslant a_n = \left(1 + \frac{1}{n}\right)^n < 3 \quad \forall n.$$

This means that the sequence has a limit which is customarily denoted as

$$\mathrm{e} = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n.$$

The number e is an irrational number and it can only be approximated: e $\simeq$ 2.7183. . . . Sometimes, in working with complicated expressions, it is convenient to use the number e as the base of logarithms. The logarithm of the number $x > 0$ to the base e is called the *natural logarithm* and denoted by $\ln x$.

---

*) The binomial theorem is given by the formula

$$(a + b)^n = a^n + \frac{n}{1!} a^{n-1} b + \frac{n(n-1)}{2!} a^{n-2} b^2 + \ldots + \frac{n(n-1)\ldots(n-k+1)}{k!} a^{n-k} b^k + \ldots + b^n,$$

where $k! = 1 \times 2 \times 3 \times \ldots \times k$.

Substituting 2 and 3 for n in the binomial formula, we obtain the formulas of abridged multiplication representing the square of the sum of two real numbers

$$(a + b)^2 = a^2 + \frac{2}{1!} ab + \frac{2 \times 1}{2!} b^2 = a^2 + 2ab + b^2$$

and the cube of the sum of two real numbers

$$(a + b)^3 = a^3 + \frac{3}{1!} a^2 b + \frac{3 \times 2}{2!} ab^2 + \frac{3 \times 2 \times 1}{3!} b^3 = a^3 + 3a^2 b + 3ab^2 + b^3,$$

which are familiar to the reader from the course of high school algebra.

Remark. To prove that the sequence $\left\{\left(1+\frac{1}{n}\right)^n\right\}$ has a limit it is helpful to apply Bernoulli's inequality. Indeed, by virtue of this inequality we have

$$a_n = \left(1+\frac{1}{n}\right)^n \geqslant 1 + n\frac{1}{n} = 2 \quad \forall n,$$

that implies that the sequence $\{a_n\}$ is bounded below.

Consider a sequence $\{b_n\}$ whose nth term b_n is given by

$$b_n = \left(1+\frac{1}{n}\right)a_n = \left(1+\frac{1}{n}\right)^{n+1} = \left(\frac{n+1}{n}\right)^{n+1}.$$

Evidently, $b_n = \left(1+\frac{1}{n}\right)a_n > 2 \quad \forall n$. Then we have

$$\begin{aligned}\frac{b_{nl}}{b_{n+1}} &= \left(\frac{n+1}{n}\right)^{n+1} \Big/ \left(\frac{n+2}{n+1}\right)^{n+2} = \frac{(n+1)^{2n+3}}{n^{n+1}(n+2)^{n+2}} \\ &= \frac{(n+1)^{2n+4}\cdot n}{(n+1)\{[(n+1)-1][(n+1)+1]\}^{n+2}} \\ &= \frac{n}{n+1}\,\frac{(n+1)^{2n+4}}{[(n+1)^2-1]^{n+2}} = \frac{n}{n+1}\left[\frac{(n+1)^2}{(n+1)^2-1}\right]^{n+2} \\ &= \frac{n}{n+1}\left[1+\frac{1}{(n+1)^2-1}\right]^{n+2}.\end{aligned}$$

Applying Bernoulli's inequality to the relation

$$\left[1+\frac{1}{(n+1)^2-1}\right]^{n+2},$$

we obtain

$$\frac{b_n}{b_{n+1}} \geqslant \frac{n}{n+1}\left(1+\frac{n+2}{(n+1)^2-1}\right) = \frac{n}{n+1}\left(1+\frac{1}{n}\right) = 1,$$

i.e., $b_n \geqslant b_{n+1}$.

Thus, the sequence $\{b_n\}$ is nonincreasing and bounded below. This means that $\{b_n\}$ has a limit and so does the sequence $\{a_n\}$ and

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} \frac{b_n}{1+\frac{1}{n}} = \lim_{n\to\infty} b_n.$$

7.3 Functions of One Variable and Limits

Notion of a function. Similarly to the concept of a set a *function* is a primary basic notion. It can be used in analysis in slightly different ways. We shall confine ourselves to the following notion of a function.

Let X be a set of real numbers x and let there be given a certain law or rule which assigns a real number y to every number x in X. Then we say that there is a function defined on X and write

$$y = f(x) \quad \text{or} \quad y = y(x), \quad x \in X^{*)}.$$

The set X is called the ***domain*** of a function and the set Y of values y specified by the function is called the ***range*** of a function. The domain is sometimes denoted by $D(f)$ and the range by $E(f)$ provided that the range of a function contains values y for every x in X.

Sometimes a function is denoted by writing f instead of $f(x)$.

A function is fully determined if there are given (i) its domain of definition X and (ii) a rule which associates every x in X with a certain value $y = f(x)$.

The functions f and g are said to be *equal* if $D(f) = D(g)$ and the identity $f(x) = g(x)$ remains true for all x in $D(f) = D(g)$. For example, the functions $y = x^2$, $-\infty < x < +\infty$, and $y = x^2$, $0 \leqslant x \leqslant 1$, are not equal since their domains are distinct; these functions are equal on the closed interval $[0, 1]$.

Examples of functions. (a) A sequence $\{a_n\}$ is a function of an integral variable whose domain is the set of natural numbers, such that $f(n) = a_n$ $(n = 1, 2, \dots)$.

(b) $y = n!$ is defined on the set of natural numbers. The relation $n!$ (read *n factorial*) is equal to the product of all integers from 1 through n, i.e.,

$$n! = 1 \times 2 \times 3 \times \dots \times n$$

and by convention $0! = 1$.

Sometimes the symbol !! is used. The relation $(2n)!!$ is assumed to be equal to the product of all even integers from 2 through $2n$. For example, $8!! = 2 \times 4 \times 6 \times 8$. The relation $(2n - 1)!!$ is equal to the product of all odd integers from 1 through $2n - 1$, e.g., $7!! = 1 \times 3 \times 5 \times 7$.

$$\text{(c) } y = \operatorname{sgn} x = \begin{cases} 1 & \text{for} \quad x > 0 \\ 0 & \text{for} \quad x = 0, \\ -1 & \text{for} \quad x < 0 \end{cases}$$

is defined at each point of the number line $-\infty < x < +\infty$. The domain

---

*) This notion describes a numeric function and can easily be generalized to the case of arbitrary sets. Let M and N be two arbitrary sets. We say that there is a function f which is defined on M and assumes values in N if every element in M is associated with one and only one element in N. Thus the function f ***maps*** the set M on the set N and is sometimes called the ***mapping***. Notice that the extended notion of a function is fully applicable to the case of sets of numbers.

of $y = \operatorname{sgn} x$ is the set containing only numbers -1, 0 and 1 (Fig. 7.6). The abbreviation $\operatorname{sgn} x$ means *signum function.*

(d) $y = [x]$, where x is a real number and $[x]$ is the largest integer not exceeding x, i.e., $[x] = n$ for $n \leqslant x < n + 1$, $n = 0, \pm 1, \pm 2, \ldots$. This function is defined at every point of the number line and its range is the set of all integers (Fig. 7.7).

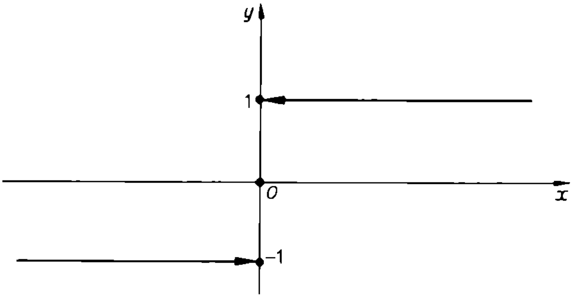


Fig. 7.6

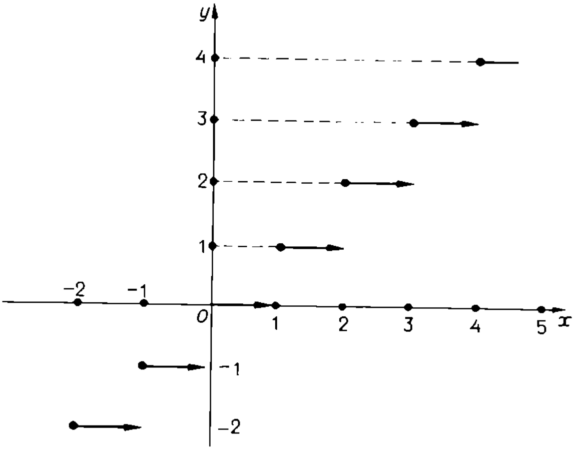


Fig. 7.7

Representations of functions. A function can be expressed by a formula, a graph or a table. Respectively, we shall speak of the *analytic, graphic* or *tabular representations* of functions.

(a) *Analytic representation.* A function $y = f(x)$ is said to be expressed *analytically* if it is defined by a formula or an equation which specifies to what operations each $x \in D(f)$ must be subjected to obtain the corresponding value y of the function. For instance, the function $y = \frac{x}{1+x^2}$ $(-\infty < x < +\infty)$ is given analytically. In this case the domain of a function, unless stated otherwise, is thought of as a set of real values x for which the formula defining the function assumes only real and finite values. In this sense the domain of a function is sometimes called the *domain of existence of a function*. For example, the domain of $y = \sqrt{1 - x^2}$ is the closed interval $-1 \leqslant x \leqslant 1$ and its range is the closed interval $0 \leqslant y \leqslant 1$. The domain of $y = \sin x$ is the infinite interval (the number line) $-\infty < x < +\infty$ and the range is the closed interval $-1 \leqslant y \leqslant 1$.

Notice that not all formulas define functions. For instance, the formula

$$y = \sqrt{1 - x^2} + \sqrt{x^2 - 4}$$

does not specify a function since for any real x at least one of the square roots does not assume a real value.

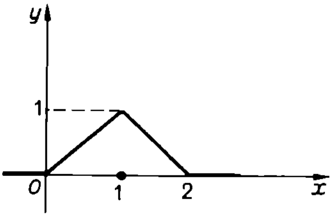


Fig. 7.8

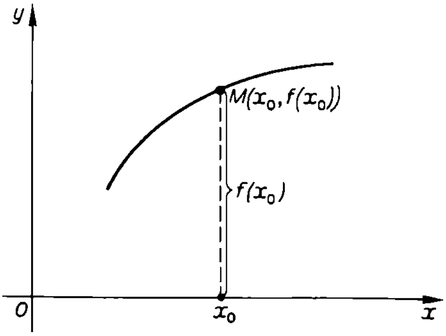


Fig. 7.9

An analytic representation does not mean that a function is always specified by a single formula. A function can be given by different relations for different parts of its domain. For example, the function $y = f(x)$ shown in Fig. 7.8 can be specified as

$$f(x) = \begin{cases} 0 & \text{for } x < 0, \\ x & \text{for } 0 \leqslant x \leqslant 1, \\ 2 - x & \text{for } 1 < x \leqslant 2, \\ 0 & \text{for } x > 2. \end{cases}$$

(b) *Graphic representation*. The graph of a function $y = f(x)$ is a locus of points $(x, f(x))$ on the xy-plane whose abscissas are the values of x and ordinates are the values of y (Fig. 7.9). We say that a function admits a graphic representation when its graph is specified.

Notice that not all functions can be given by graphs. For example, the Dirichlet function

$$D(x) = \begin{cases} 1 & \text{for rational } x, \\ 0 & \text{for irrational } x \end{cases}$$

does not admit a graphic representation. The domain of this function is the number line and the range comprises two numbers 0 and 1.

(c) *Tabular representation.* A function is said to be specified in a tabular form if it is represented by means of a table which contains numerical values of the variable x and the corresponding values of y.

Limit of a function at a point. A limit is a basic concept of mathematical analysis.

Cauchy criterion for limits. Let $f(x)$ be a function defined in a neighbourhood Ω of a point x_0, except probably at x_0. Then a number A is a *limit* of $f(x)$ at x_0 if, given any (whatever small) number $\varepsilon > 0$, there exists a number $\delta > 0$ such that

$$|f(x) - A| < \varepsilon,$$

whenever $|x - x_0| < \delta$ and $x \neq x_0$.

To signify that the number A is a limit of $f(x)$ at the point x_0 we write

$$\lim_{x \to x_0} f(x) = A.$$

Using logical symbols we can express this criterion as

$$\left(\lim_{x \to x_0} f(x) = A\right) \Leftrightarrow \forall \varepsilon > 0 \quad \exists \delta > 0 \quad \forall x, \quad x \neq x_0,$$

$$|x - x_0| < \delta \Rightarrow |f(x) - A| < \varepsilon.$$

Example. Let $f(x) = 2x + 3$ and $x_0 = 1$. Verify that $\lim_{x \to 1} f(x) = 5$.

◄ Clearly, $f(x)$ is defined everywhere including the point $x_0 = 1$ where $f(1) = 5$. Let $\varepsilon > 0$ be an arbitrary number. For the inequality $|(2x + 3) - 5| < \varepsilon$ to hold it is necessary that

$$|2x - 2| < \varepsilon \Rightarrow 2|x - 1| < \varepsilon \Rightarrow |x - 1| < \frac{\varepsilon}{2}.$$

Thus, if $\delta = \frac{\varepsilon}{2}$, we have $0 < |x - 1| < \delta = \frac{\varepsilon}{2}$ and $|f(x) - 5| < \varepsilon$. Whence it follows that the number 5 is the limit of $f(x) = 2x + 3$ at the point $x_0 = 1$. ►

Geometric interpretation of a limit at a point. Let $f(x)$ be given by a graph so that the values of $f(x)$ are equal to the ordinates of the curve

M_1M for $x < x_0$ and to the ordinates of the curve MM_2 for $x > x_0$ (Fig. 7.10). And let the value $f(x_0)$ be equal to the ordinate of the point N. In other words, we assume that the graph of $f(x)$ is obtained from the "good" curve M_1MM_2 by replacing N for M.

Let us show that at x_0 the function $f(x)$ has a limit equal to the ordinate A of M.

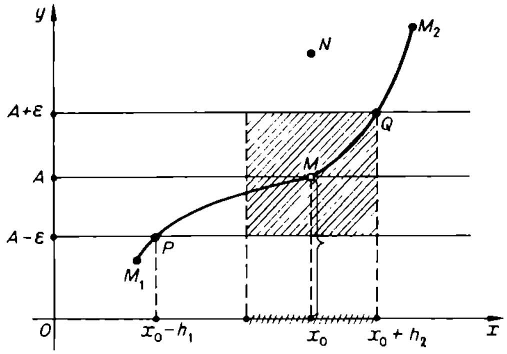


Fig. 7.10

◄ We choose any (whatever small) $\varepsilon > 0$ and fix the points with the ordinates A, $A - \varepsilon$, $A + \varepsilon$ on the y-axis. Let P and Q be the points of intersection of the graph of $y = f(x)$ with the lines $y = A - \varepsilon$ and $y = A + \varepsilon$ and let $x_0 - h_1$ and $x_0 + h_2$, where $h_1 > 0$ and $h_2 > 0$, be the abscissas of P and Q, respectively. We easily see for any $x \neq x_0$ in the interval $(x_0 - h_1, x_0 + h_2)$ the value of $f(x)$ lies between $A - \varepsilon$ and $A + \varepsilon$, i.e., for all $x \neq x_0$ we have

$$A - \varepsilon < f(x) < A + \varepsilon$$

whenever $x_0 - h_1 < x < x_0 + h_2$.

Let $\delta = \min\{h_1, h_2\}$ be the smallest of h_1 and h_2. Then the interval $(x_0 - \delta, x_0 + \delta)$ is contained in the interval $(x_0 - h_1, x_0 + h_2)$. Hence, the inequality $A - \varepsilon < f(x) < A + \varepsilon$ or, equivalently,

$$|f(x) - A| < \varepsilon$$

remains true for all $x \neq x_0$ belonging to the interval $(x_0 - \delta, x_0 + \delta)$, i.e., for all x satisfying the condition

$$0 < |x - x_0| < \delta.$$

Whence it follows that $\lim\limits_{x \to x_0} f(x) = A$. ►

Thus we infer that the function $y = f(x)$ has a limit A at a point x_0 provided that for any (whatever narrow) strip between the lines $y = A - \varepsilon$ and $y = A + \varepsilon$ there exists $\delta > 0$ such that the graph of $y = f(x)$ lies inside this strip for all x belonging to the deleted δ-neighourhood of x_0.

Remarks. (1) In general the value of δ depends on the value of ε, i.e., $\delta = \delta(\varepsilon)$.

(2) When we deal with a limit of a function at a point x_0 we exclude x_0 from consideration. Thus a limit of f at x_0 is independent of the value of f at x_0. Furthermore, a function can even be undefined at x_0. So any two functions equal to each other in a neighbourhood of x_0, except probably at the point x_0, where they can differ or even can both be undefined, have the same limit as $x \to x_0$ or have no limits at all. In particular, this implies that when computing the limit of a fraction at x_0 we may reduce this fraction by like factors vanishing at $x = x_0$.

Examples. **(1)** Compute $\lim\limits_{x \to 0} \dfrac{x}{x}$.

◀ Notice that $f(x) = \dfrac{x}{x} = 1$ for all $x \neq 0$ and is undefined at $x = 0$. By the definition of a limit the point $x = 0$ is excluded from consideration. So

$$\lim_{x \to 0} \frac{x}{x} = \lim_{x \to 0} 1 = 1. \quad ▶$$

(2) Compute $\lim\limits_{x \to 0} f(x)$ (see Fig. 7.11), where

$$f(x) = \begin{cases} x^2 & \text{for } x \neq 0, \\ 1 & \text{for } x = 0. \end{cases}$$

◀ The function $g(x) = x^2$, $-\infty < x < +\infty$, is equal to $f(x)$ everywhere, except at $x = 0$, and its limit at $x = 0$ is equal to zero, i.e., $\lim\limits_{x \to 0} g(x) = 0$ (show this!). Hence, $\lim\limits_{x \to 0} f(x) = 0$. ▶

Problems. Write down the Cauchy criterion for the following limits

(1) $\lim\limits_{x \to 1} f(x) = 5$, **(2)** $\lim\limits_{x \to 0} f(x) = 0$,

(3) $\lim\limits_{x \to -2} f(x) = 1$, **(4)** $\lim\limits_{x \to -3} f(x) = -2$.

Sequential criterion for limits. Let $f(x)$ be a function defined in a neighbourhood Ω of x_0, except probably at x_0. And let $\{x_n\}$, where $x_n \in \Omega$ and $x_n \neq x_0$, be a sequence of values of x that converges to x_0. Then a number A is the *limit* of $f(x)$ provided that for any $\{x_n\}$ the corresponding sequence $\{f(x_n)\}$ of the values of $f(x)$ converges to A.

This criterion is convenient to apply when we wish to determine whether the function $f(x)$ has a limit at x_0 or not. It suffices to define a sequence

$\{f(x_n)\}$ which has no limit or two sequences $\{f(x_n)\}$ and $\{f(x_n')\}$ having distinct limits.

◀ Consider the function $f(x) = \sin\frac{1}{x}$ defined everywhere, except at the point $x = 0$ (Fig. 7.12).

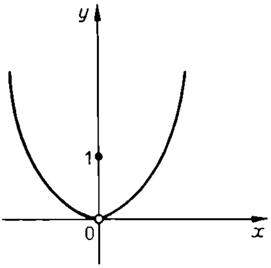


Fig. 7.11

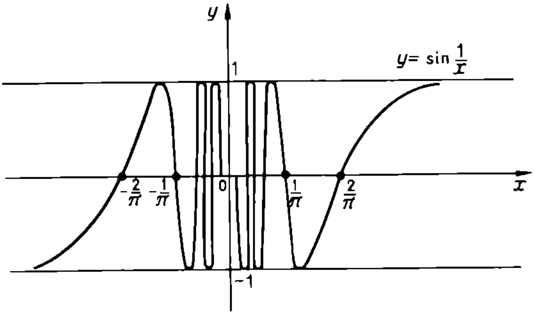


Fig. 7.12

We have

$$\begin{cases} x_n = \dfrac{1}{n\pi} \xrightarrow[n\to\infty]{} 0, \\ f(x_n) = \sin n\pi = 0, \end{cases} \quad \text{and} \quad \begin{cases} x_n' = \dfrac{1}{\dfrac{\pi}{2} + 2n\pi} \xrightarrow[n\to\infty]{} 0, \\ f(x_n') = \sin\left(\dfrac{\pi}{2} + 2n\pi\right) = 1. \end{cases}$$

Hence we have found two sequences $\{x_n\}$ and $\{x_n'\}$ converging to $x = 0$, for which the corresponding sequences of the values of $f(x)$ converge to different limits, viz., $\{f(x_n)\}$ converges to zero while $\{f(x_n')\}$ converges to 1. By the sequential criterion this means that $f(x) = \sin\frac{1}{x}$ has no limit at $x = 0$. ▶

Remark. It can be proved that both criteria for limits at a point are equivalent.

Limit theorems. We shall derive some basic results which are of importance in examining limits of functions of one variable.

Theorem 7.10. *Let $f(x)$ have a limit at a point x_0. Then this limit is unique.*

◀ Let $\lim\limits_{x\to x_0} f(x) = A$. We shall show that there exists no number $B \neq A$ which is a limit of $f(x)$ at x_0.

The statement $\lim\limits_{x\to x_0} f(x) \neq B$ can be written as

$$\exists\varepsilon > 0 \quad \forall\delta > 0 \quad \exists x, \quad x \neq x_0, \quad |x - x_0| < \delta \wedge |f(x) - B| \geqslant \varepsilon.$$

Using $\big||a| - |b|\big| \leqslant |a - b|$, we have

$$|f(x) - B| = |(f(x) - A) - (B - A)| \geqslant \big||f(x) - A| - |B - A|\big|$$
$$= \big||B - A| - |f(x) - A|\big|. \qquad (*)$$

Let $\varepsilon = \dfrac{|B - A|}{2} > 0$. Since $\lim\limits_{x \to x_0} f(x) = A$, then for the chosen $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|f(x) - A| < \varepsilon \quad \forall x, \quad x \neq x_0, \quad |x - x_0| < \delta.$$

Applying $(*)$, we reduce the above statement to the following

$$|f(x) - B| \geqslant \varepsilon \quad \forall x, \quad x \neq x_0, \quad |x - x_0| < \delta.$$

Hence, we have found $\varepsilon > 0$ such that for any (whatever small) $\delta > 0$ there exists a point $x \neq x_0$ for which $0 < |x - x_0| < \delta$ and $|f(x) - B| \geqslant \varepsilon$. Whence it follows that $B \neq \lim\limits_{x \to x_0} f(x)$. ►

A function $f(x)$ is said to be *bounded* in a neighbourhood of a point x_0 if there exist numbers $M > 0$ and $\delta > 0$ such that

$$|f(x)| \leqslant M \quad \forall x \in (x_0 - \delta, \quad x_0 + \delta),$$

where $f(x)$ is defined.

Theorem 7.11. *Let $f(x)$ have a finite limit at x_0. Then $f(x)$ is bounded in a neighbourhood of x_0, i.e., $\exists M > 0$, $\exists \delta > 0$ and*

$$|f(x)| \leqslant M \quad \forall x \in (x_0 - \delta, \quad x_0 + \delta), \quad x \neq x_0.$$

◄ Let $\lim\limits_{x \to x_0} f(x) = A$. Then for any $\varepsilon > 0$, say, $\varepsilon = 1$, there exists $\delta > 0$ such that

$$|f(x) - A| < 1$$

whenever $x \neq x_0$ and $|x - x_0| < \delta$.

Observe that

$$|f(x)| - |A| \leqslant |f(x) - A|$$

is always true. Then we have

$$|f(x)| < |A| + 1.$$

Let $M = |A| + 1$ if $f(x)$ is undefined at x_0 and let $M = \max\{|A| + 1, |f(x_0)|\}$ if $f(x)$ is defined at x_0. Then

$$|f(x)| \leqslant M$$

holds at every point x in the interval $(x_0 - \delta, x_0 + \delta)$. By definition this means that $f(x)$ is bounded in a neighbourhood of x_0. ►

Theorem 7.11 implies that every function having a limit is bounded. However, the converse is not always true, i.e., a function can be bounded in a neighbourhood of x_0 but have no limit at x_0. For example, $f(x) = \sin\frac{1}{x}$ is bounded in a neighbourhood of $x = 0$ since $\left|\sin\frac{1}{x}\right| \leqslant 1 \ \forall x, x \neq 0$ but $f(x)$ has no limit at $x = 0$.

The following two theorems are easy to interpret geometrically.

Theorem 7.12. *Let $f(x) \leqslant \varphi(x)$ for all x in a neighbourhood of x_0, except probably at x_0, and let $f(x)$ and $\varphi(x)$ have limits at x_0. Then* (see Fig. 7.13)

$$\lim_{x \to x_0} f(x) \leqslant \lim_{x \to x_0} \varphi(x).$$

Notice that in general from the strict inequality $f(x) < \varphi(x)$ it follows that

$$\lim_{x \to x_0} f(x) \leqslant \lim_{x \to x_0} \varphi(x).$$

For example, let

$$f(x) = x^2 \quad \text{and} \quad \varphi(x) = \begin{cases} 2x^2 & \text{for } x \neq 0, \\ 1 & \text{for } x = 0. \end{cases}$$

Obviously, we have $f(x) < \varphi(x) \quad \forall x$ whereas

$$\lim_{x \to 0} f(x) = \lim_{x \to 0} \varphi(x) = 0.$$

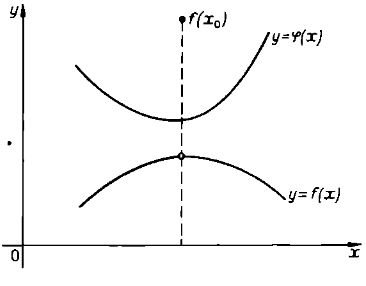


Fig. 7.13

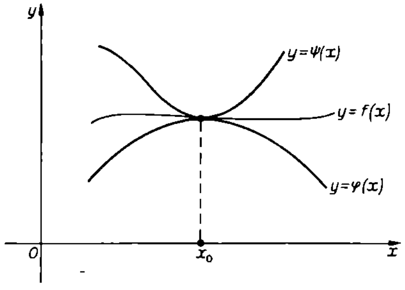


Fig. 7.14

Theorem 7.13. *Let $\varphi(x) \leqslant f(x) \leqslant \psi(x)$ be true for all x in a neighbourhood of x_0, except probably at x_0, and let $\varphi(x)$ and $\psi(x)$ have the same limit A at x_0. Then $f(x)$ also has the limit A at x_0* (Fig. 7.14).

Limit of a function when the variable tends to infinity. Let $f(x)$ be defined at every point of the number line or at all x whenever $|x| > K$ for some $K > 0$ so that we can compute the values of $f(x)$ for whatever large magnitudes of x.

Definition. We say that a number A is a limit of $f(x)$ as x tends to infinity and write

$$\lim_{x \to \infty} f(x) = A$$

provided that for any $\varepsilon > 0$ there exists a number $N > 0$ such that

$$|f(x) - A| < \varepsilon,$$

whenever $|x| > N$.

Substituting $x > N$ and $x < -N$ for $|x| > N$, we have $A = \lim\limits_{x \to +\infty} f(x)$ and $A = \lim\limits_{x \to -\infty} f(x)$, respectively. Whence it follows that

$$A = \lim_{x \to \infty} f(x)$$

if and only if both $A = \lim\limits_{x \to +\infty} f(x)$ and $A = \lim\limits_{x \to -\infty} f(x)$ hold.

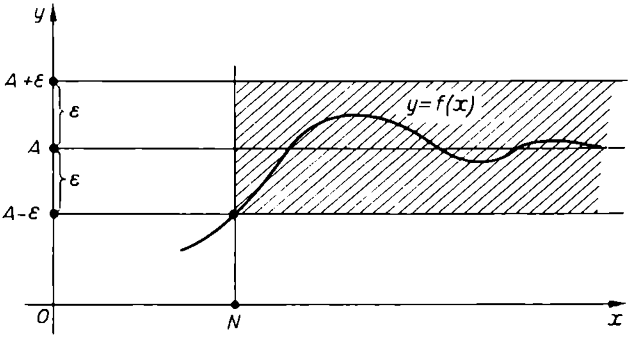


Fig. 7.15

Geometrically the statement $A = \lim\limits_{x \to +\infty} f(x)$ means that, given any (whatever narrow) strip between the lines $y = A - \varepsilon$ and $y = A + \varepsilon$, there exists a line $x = N > 0$ such that for all $x > N > 0$ the graph of $y = f(x)$ is contained inside this strip (Fig. 7.15). In this case we say that the curve $y = f(x)$ asymptotically tends to the line $y = A$ as $x \to +\infty$.

Example. Consider the function $f(x) = \dfrac{1}{x^2 + 1}$ defined at all points of the number line in which case the nominator is bounded everywhere and the denominator increases to infinity as $|x| \to +\infty$. This prompts us to assume that $\lim\limits_{x \to \infty} f(x) = 0$.

◄ Let us choose $\varepsilon > 0$ such that $0 < \varepsilon \leqslant 1$. For the inequality

$$\left| \frac{1}{x^2+1} - 0 \right| < \varepsilon$$

to hold there must hold $\dfrac{1}{x^2+1} < \varepsilon$. Whence $x^2 + 1 > \dfrac{1}{\varepsilon}$ and $|x| > \sqrt{\dfrac{1}{\varepsilon} - 1}$. Thus, if we choose $N = \sqrt{\dfrac{1}{\varepsilon} - 1}$ then $\left| \dfrac{1}{x^2+1} - 0 \right| < \varepsilon$ whenever $|x| > N$, i.e., $A = 0$ is the limit of the given function as $x \to \infty$.

Notice that the inequality $\dfrac{1}{\varepsilon} - 1 \geqslant 0$ is true only for $\varepsilon \leqslant 1$. For $\varepsilon > 1$ it immediately follows that $\dfrac{1}{x^2+1} < \varepsilon$ for all $x \in \mathbb{R}$.

The graph of an even function $y = \dfrac{1}{x^2+1}$ asymptotically tends to the line $y = 0$ as $x \to \pm\infty$. ►

Problems. We leave it to the reader to interpret the following statements in terms of inequalities involved in the definition of a limit

(1) $\lim\limits_{x\to\infty} f(x) = -3$; (2) $\lim\limits_{x\to+\infty} f(x) = 1$; (3) $\lim\limits_{x\to-\infty} f(x) = 0$.

7.4 Infinitesimals and Infinities

Notion of an infinitesimal. Let a function $\alpha(x)$ be defined in a neighbourhood of a point x_0, except probably at x_0. Then $\alpha(x)$ is said to be an *infinitely small function* or an *infinitesimal* as x tends to x_0 if $\alpha(x)$ has at x_0 a limit equal to zero, i.e.,

$$\lim_{x\to x_0} \alpha(x) = 0.$$

For instance, the function $\alpha(x) = x - 1$ shown in Fig. 7.16 is an infinitesimal as $x \to 1$ since $\lim\limits_{x\to 1} (x - 1) = 0$.

In general the function $\alpha(x) = x - x_0$ offers an elementary example of an infinitesimal as $x \to x_0$.

Applying the Cauchy criterion for limits we can define an infinitesimal as follows.

A function $\alpha(x)$ is an infinitesimal as $x \to x_0$ if given any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|\alpha(x)| < \varepsilon$$

whenever $|x - x_0| < \delta$ and $x \neq x_0$.

Using logical symbols we write

$$(\alpha(x) \text{ is an infinitesimal as } x \to x_0) \Leftrightarrow \forall \varepsilon > 0 \quad \exists \delta > 0 \quad \forall x,\ x \neq x_0,$$
$$|x - x_0| < \delta \Rightarrow |\alpha(x)| < \varepsilon.$$

Analogously, we can define infinitesimals as $x \to \infty$, $x \to +\infty$ and $x \to -\infty$.

We say that $\alpha(x)$ is an infinitesimal as $x \to \infty$ if $\lim\limits_{x \to \infty} \alpha(x) = 0$.

If $\lim\limits_{x \to +\infty} \alpha(x) = 0$ or $\lim\limits_{x \to -\infty} \alpha(x) = 0$ the function $\alpha(x)$ is said to be an infinitesimal as $x \to +\infty$ or as $x \to -\infty$, respectively.

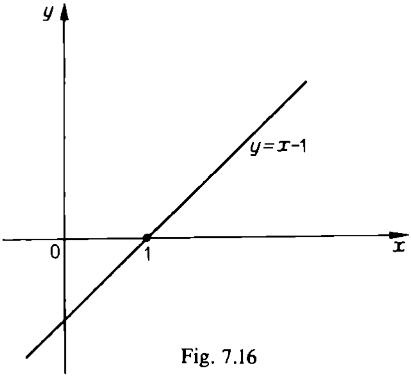


Fig. 7.16

For instance, the function $\alpha(x) = \frac{1}{x}$, $x \neq 0$, is an infinitesimal as $x \to \infty$, since $\lim\limits_{x \to \infty} \frac{1}{x} = 0$. The function $\alpha(x) = e^{-x}$ is an infinitesimal as $x \to +\infty$, for $\lim\limits_{x \to +\infty} e^{-x} = 0$.

In what follows we shall confine ourselves to infinitesimals as x tends to x_0 by leaving to the reader a generalization of theorems and concepts to the cases when $x \to \infty$, $x \to +\infty$ and $x \to -\infty$.

Properties of infinitesimals. Now we shall prove some important theorems which enable us to operate on infinitesimals.

Theorem 7.14. *If $\alpha(x)$ and $\beta(x)$ are infinitesimals as $x \to x_0$ so is the sum $\alpha(x) + \beta(x)$ as $x \to x_0$.*

◄ Since $\alpha(x)$ is an infinitesimal as $x \to x_0$ then for any $\varepsilon > 0$, there exists $\delta_1 > 0$ such that

$$|\alpha(x)| < \frac{\varepsilon}{2}, \tag{*}$$

whenever $|x - x_0| < \delta_1$ and $x \neq x_0$.

Analogously, for $\beta(x)$ we have $\delta_2 > 0$ such that

$$|\beta(x)| < \frac{\varepsilon}{2} \tag{**}$$

whenever $|x - x_0| < \delta_2$ and $x \neq x_0$.

Let $\delta = \min\{\delta_1, \delta_2\}$. Then both (*) and (**) hold whenever $|x - x_0| < \delta$ and $x \neq x_0$. Hence

$$|\alpha(x) + \beta(x)| \leqslant |\alpha(x)| + |\beta(x)| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

$$\forall x, \quad x \neq x_0, \quad |x - x_0| < \delta.$$

Thus the sum $\alpha(x) + \beta(x)$ is an infinitesimal as $x \to x_0$. ►

Similar theorem can easily be proved for any finite number of infinitesimals as $x \to x_0$.

Theorem 7.15. *If a funciton $\alpha(x)$ is an infinitesimal as $x \to x_0$ and if a function $f(x)$ is bounded in a neighbourhood of x_0 then the product $\alpha(x)f(x)$ is an infinitesimal as $x \to x_0$.*

◄ Since $f(x)$ is bounded in a neighbourhood of x_0 there exist numbers $\delta_1 > 0$ and $M > 0$ such that

$$|f(x)| \leqslant M$$

for all $x \in (x_0 - \delta_1, x_0 + \delta_1)$ where $f(x)$ is defined.

Consider an arbitrary $\varepsilon > 0$. For an infinitesimal $\alpha(x)$ there exists $\delta_2 > 0$ such that, as $x \to x_0$,

$$|\alpha(x)| < \frac{\varepsilon}{M}$$

whenever $|x - x_0| < \delta_2$ and $x \neq x_0$.

Let $\delta = \min\{\delta_1, \delta_2\}$. Then both

$$|f(x)| \leqslant M \quad \text{and} \quad |\alpha(x)| < \frac{\varepsilon}{M}$$

are true whenever $|x - x_0| < \delta$ and $x \neq x_0$.

Hence

$$|\alpha(x)f(x)| = |\alpha(x)| \cdot |f(x)| < \frac{\varepsilon}{M} \cdot M = \varepsilon \quad \forall x, \quad x \neq x_0, \quad |x - x_0| < \delta.$$

By definition we infer that the product $\alpha(x)f(x)$ is an infinitesimal as $x \to x_0$. ►

Example. Consider the function $y = x \sin \frac{1}{x}$ (Fig. 7.17).

The relation $x \sin \frac{1}{x}$ can be considered as the product of the functions $\alpha(x) = x$ and $f(x) = \sin \frac{1}{x}$. The function $\alpha(x)$ is an infinitesimal as $x \to 0$ and $f(x) = \sin \frac{1}{x}$ is defined everywhere, except at the point $x = 0$, and is bounded in any neighbourhood of $x = 0$. Then by Theorem 7.15 $y = x \sin 1/x$ is an infinitesimal as $x \to 0$ so that

$$\lim_{x \to 0} x \sin \frac{1}{x} = 0.$$

Corollary. If a function $\alpha(x)$ is an infinitesimal as $x \to x_0$ and if a function $f(x)$ has a finite limit at x_0 then the product $\alpha(x)f(x)$ is an infinitesimal as $x \to x_0$.

◄ To prove this corollary it suffices to observe that a function having a limit at x_0 is bounded in a neighbourhood of x_0. ►

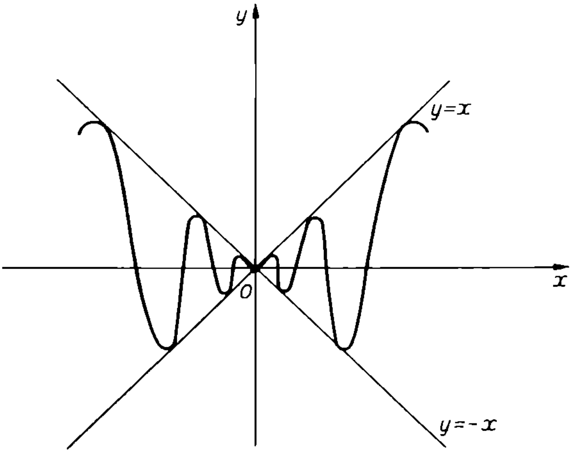


Fig. 7.17

Lemma. If a function $f(x)$ has a nonzero limit at a point x_0 then a function $\frac{1}{f(x)}$ is bounded in a deleted neighbourhood of x_0.

◄ Let $\lim_{x \to x_0} f(x) = A \neq 0$. Then for any $\varepsilon > 0$, say $\varepsilon = \frac{|A|}{2}$, there exists $\delta > 0$ such that for all x

$$|A - f(x)| < \frac{|A|}{2}$$

whenever $0 < |x - x_0| < \delta$.

Since the inequality $|A - f(x)| \geqslant |A| - |f(x)|$ is always true we have

$$|A| - |f(x)| < \frac{|A|}{2}.$$

Whence

$$|f(x)| > \frac{|A|}{2} \quad \forall x, \quad x \neq x_0, \quad |x - x_0| < \delta.$$

This means that a function $\dfrac{1}{f(x)}$ is defined for all x, $x \neq x_0$, $|x - x_0| < \delta$ and

$$\left|\frac{1}{f(x)}\right| = \frac{1}{|f(x)|} < \frac{2}{|A|}.$$

Hence the function $\dfrac{1}{f(x)}$ is bounded in a deleted neighbourhood of x_0. ►

Theorem 7.16. *If* $\alpha(x)$ *is an infinitesimal as* $x \to x_0$ *and if* $f(x)$ *has a nonzero limit at* x_0 *then the quotient* $\dfrac{\alpha(x)}{f(x)}$ *is an infinitesimal as* $x \to x_0$.

◄ Express $\dfrac{\alpha(x)}{f(x)}$ as

$$\frac{\alpha(x)}{f(x)} = \alpha(x)\frac{1}{f(x)}.$$

By virtue of the above lemma the function $\dfrac{1}{f(x)}$ is bounded in a neighbourhood of a point x_0. Then, applying Theorem 7.15, we infer that the function $\alpha(x)\dfrac{1}{f(x)}$ is an infinitesimal as $x \to x_0$. ►

The condition $\lim\limits_{x \to x_0} f(x) \neq 0$ is essential when we wish to apply Theorem 7.16. For example, let $\alpha(x) = x$ and $f(x) = x^2$. The function $\alpha(x)$ is an infinitesimal as $x \to 0$ and the function $f(x)$ has a limit equal to zero at $x = 0$, i.e., $f(x)$ is an infinitesimal as $x \to 0$. Evidently, the quotient $\dfrac{\alpha(x)}{f(x)} = \dfrac{x}{x^2} = \dfrac{1}{x}$, $x \neq 0$, is not an infinitesimal as $x \to 0$. In general a quotient of two infinitesimals is not an infinitesimal.

Notion of infinity. Let a function $f(x)$ be defined in a neighbourhood of a point x_0, except probably at x_0. Then a function $f(x)$ is said to be an *infinitely large function* or an *infinity* as x tends to x_0 if, given any (whatever large) number M, there exists a number $\delta > 0$ such that

$$|f(x)| > M$$

whenever $|x - x_0| < \delta$ and $x \neq x_0$. In this case we write

$$\lim_{x \to x_0} f(x) = \infty.$$

We say that $f(x)$ has an infinite limit as $x \to x_0$. We also say that a limit of $f(x)$ is equal to A with the understanding that A is any (whatever large) finite number.

Using logical symbols we can write

$$(f(x) \text{ is an infinity as } x \to x_0) \Leftrightarrow \forall M > 0 \quad \exists \delta > 0, \quad \forall x, \quad x \neq x_0 \quad |x - x_0| < \delta \Rightarrow |f(x)| > M.$$

Substituting $f(x) > M$ or $f(x) < -M$ for $|f(x)| > M$, we have

$$\lim_{x \to x_0} f(x) = +\infty \quad \text{or} \quad \lim_{x \to x_0} f(x) = -\infty,$$

respectively.

For example, the function $f(x) = \frac{1}{x}$ defined at every $x \neq 0$ is an infinity as $x \to 0$ (Fig. 7.18). The function $f(x) = \frac{1}{x^2}$ defined at every $x \neq 0$ is an infinity tending to $+\infty$ as $x \to 0$ (Fig. 7.19).

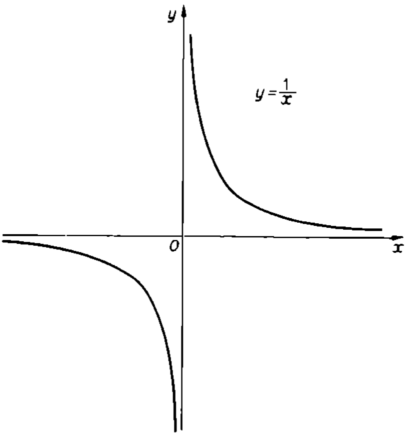


Fig. 7.18

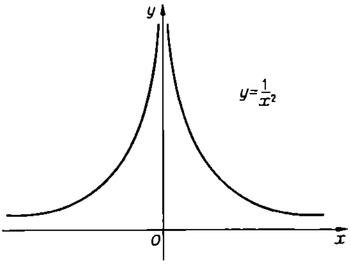


Fig. 7.19

Let us show that $f(x) = \frac{1}{x}$ is an infinity as $x \to 0$. Consider any whatever large $M > 0$. For the inequality $|f(x)| = \left|\frac{1}{x}\right| = \frac{1}{|x|} > M$ to hold it is necesary and sufficient that the inequality $|x| = |x - 0| < \frac{1}{M}$ hold. So if we put $\delta = \frac{1}{M}$, then $|f(x)| = \frac{1}{|x|} > M$ whenever $|x - 0| = |x| < \frac{1}{M}$ and $x \neq 0$. Whence we infer that $f(x) = \frac{1}{x}$ is an infinity as $x \to 0$.

To interpret the notion of infinity geometrically look at Fig. 7.20. A function $f(x)$ is an infinity as $x \to x_0$ if given any (whatever wide) horizontal strip between the lines $y = -M$ and $y = M$ there exist two vertical lines $x = x_0 - \delta$ and $x = x_0 + \delta$ such that the graph of $y = f(x)$, $x \neq x_0$, lies outside this horizontal strip whenever x is contained in the open interval $(x_0 - \delta, x_0 + \delta)$. Notice that not every function $f(x)$ which is unbounded in a neighbourhood of x_0 is an infinite function as $x \to x_0$. For instance, the function $f(x) = \frac{1}{x}\sin\frac{1}{x}$ is unbounded in a neighbourhood of $x = 0$; however, $f(x)$ is not an infinity as $x \to 0$. (As an exercise draw a figure to interpret this example geometrically.)

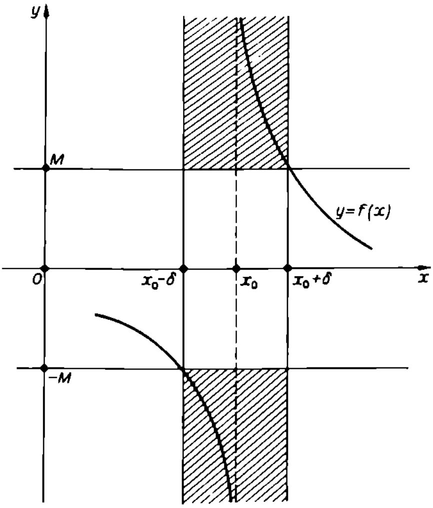


Fig. 7.20

We say that a function $f(x)$ is an infinity as $x \to \infty$ and write

$$\lim_{x \to \infty} f(x) = \infty$$

if, given any (whatever large) number $M > 0$, there exists a number $N > 0$ such that

$$|f(x)| > M$$

whenever $|x| > N$.

Example. The function $f(x) = x$ is an infinity as $x \to \infty$.

◀ Indeed, $\forall M > 0\ \exists N > 0$, say $N = M$, such that $|f(x)| = |x| > M$ whenever $|x| > N$, $\forall x$. ▶

Analogously, we can define infinities as $x \to +\infty$ and as $x \to -\infty$.

Relations between infinitesimals and infinities. The following theorems lay down the relations between infinitesimals and infinities.

Theorem 7.17. *If a function $f(x)$ is an infinity as $x \to x_0$ then a function $\alpha(x) = \dfrac{1}{f(x)}$ is an infinitesimal as $x \to x_0$.*

◀ Let $\varepsilon > 0$ be an arbitrary (whatever small), number. Since $f(x)$ is an infinity as $x \to x_0$ then for any $M > 0$, say, $M = \dfrac{1}{\varepsilon}$, there exists $\delta > 0$ such that

$$|f(x)| > M = \frac{1}{\varepsilon}$$

whenever $|x - x_0| < \delta$ and $x \neq x_0$, in which case there will be defined a function $\alpha(x) = \dfrac{1}{f(x)}$ and

$$|\alpha(x)| = \frac{1}{|f(x)|} < \frac{1}{M} = \varepsilon.$$

Thus

$$\forall \varepsilon > 0\ \exists \delta > 0,\ \forall x,\ x \neq x_0,\ |x - x_0| < \delta \Rightarrow |\alpha(x)| < \varepsilon.$$

Whence we infer that $\alpha(x) = \dfrac{1}{f(x)}$ is an infinity as $x \to x_0$. ▶

Similarly we can prove the following theorem.

Theorem 7.18. *If a function $\alpha(x)$ is an infinitesimal as $x \to x_0$ and if $\alpha(x)$ is distinct from zero in the neighbourhood $(x_0 - \delta, x_0 + \delta)$ of x_0, except probably at x_0, then the function $f(x) = \dfrac{1}{\alpha(x)}$ is an infinity as $x \to x_0$.*

Problems. Using inequalities write down the definitions for the limits:

(1) $\lim\limits_{x \to 1} f(x) = \infty$, **(2)** $\lim\limits_{x \to -3} f(x) = +\infty$,

(3) $\lim\limits_{x \to 0} f(x) = -\infty$, **(4)** $\lim\limits_{x \to +\infty} f(x) = +\infty$,

(5) $\lim\limits_{x \to \infty} f(x) = +\infty$, **(6)** $\lim\limits_{x \to -\infty} f(x) = -\infty$.

In conclusion of this section we shall consider the rational function

$$y(x) = \frac{a_0x^m + a_1x^{m-1} + \ldots + a_m}{b_0x^n + b_1x^{n-1} + \ldots + b_n} \quad (a_0 \neq 0,\ b_0 \neq 0),$$

which is a quotient of two polynomials in x of degrees m and n, respectively. For sufficiently large $|x|$ the denominator is distinct from zero; thus the quotient makes sense. Then dividing the nominator and denominator by x^n, we obtain

$$y(x) = \frac{a_0x^{m-n} + a_1x^{m-n-1} + \ldots + a_mx^{-n}}{b_0 + b_1x^{-1} + \ldots + b_nx^{-n}}.$$

It is easily seen that a limit of the denominator as $x \to \infty$ is equal to $b_0 \neq 0$. The nominator increases infinitely if $m > n$, has a limit equal to a_0 if $m = n$ or has a limit equal to zero if $m < n$. Therefore, the behaviour of this function as $x \to \infty$ is fully determined by the formulas

$$\lim_{x\to\infty} \frac{a_0x^m + a_1x^{m-1} + \ldots + a_m}{b_0x^n + b_1x^{n-1} + \ldots + b_n} = \begin{cases} \infty, & m > n, \\ \dfrac{a_0}{b_0}, & m = n, \\ 0, & m < n. \end{cases}$$

7.5 Operations on Limits

Relationship between a function having a limit and an infinitesimal. We shall prove a theorem which will be frequently used below.

Theorem 7.19. *Let a function $f(x)$ be defined in a neighbourhood of a point x_0, except probably at x_0. For $f(x)$ to have a limit A at x_0 it is necessary and sufficient that $f(x)$ admit a representation of the form*

$$f(x) = A + \alpha(x),$$

where $\alpha(x)$ is an infinitesimal as $x \to x_0$.

◄ *Necessity.* Let $f(x)$ have a limit equal to A at x_0, i.e.,

$$\lim_{x\to x_0} f(x) = A.$$

We put

$$\alpha(x) = f(x) - A \qquad (*)$$

or, equivalently,

$$f(x) = A + \alpha(x)$$

and prove that $\alpha(x)$ is an infinitesimal as $x \to x_0$.

Let $\varepsilon > 0$ be an arbitrary number. Since $\lim\limits_{x\to x_0} f(x) = A$, then for the given $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|f(x) - A| < \varepsilon$$

whenever $|x - x_0| < \delta$ and $x \neq x_0$.

Taking into account (*), we can write

$$|\alpha(x)| < \varepsilon$$

or

$$\forall\varepsilon > 0\ \exists\delta > 0,\ \forall x,\ x \neq x_0,\ |x - x_0| < \delta \Rightarrow |\alpha(x)| < \varepsilon.$$

Whence it follows that $\alpha(x)$ is an infinitesimal as $x \to x_0$.

Sufficiency. Let $f(x)$ admit a representation of the form

$$f(x) = A + \alpha(x), \qquad (**)$$

where A is a constant quantity and $\alpha(x)$ is an infinitesimal as $x \to x_0$. We shall prove that $f(x)$ has a limit equal to A at x_0.

Suppose that $\varepsilon > 0$ is an arbitrary number. Since $\alpha(x)$ is an infinitesimal as $x \to x_0$ there exists $\delta > 0$ such that

$$|\alpha(x)| < \varepsilon$$

whenever $|x - x_0| < \delta$ and $x \neq x_0$.

But by virtue of (**) we have $\alpha(x) = f(x) - A$ and $|f(x) - A| < \varepsilon$. Thus

$$\forall\varepsilon > 0,\ \exists\delta > 0,\ \forall x,\ x \neq x_0,\ |x - x_0| < \delta \Rightarrow |f(x) - A| < \varepsilon.$$

This means that $A = \lim\limits_{x \to x_0} f(x)$. ►

Example. Let $f(x) = x$ and $x_0 = 2$. Then $\lim\limits_{x \to 2} f(x) = 2$ and

$$f(x) = 2 + (x - 2)$$

where 2 is a constant quantity, $x - 2$ is an infinitesimal as $x \to 2$, i.e., $\alpha(x) = x - 2$.

Arithmetical operations on limits. Let $f(x)$ and $\varphi(x)$ be defined in a neighbourhood of a point x_0, except probably at x_0.

Theorem 7.20. *If* $\lim\limits_{x \to x_0} f(x) = A$ *and* $\lim\limits_{x \to x_0} \varphi(x) = B$ *then*

(a) $\lim\limits_{x \to x_0} [f(x) \pm \varphi(x)] = A \pm B,$

(b) $\lim\limits_{x \to x_0} [f(x) \cdot \varphi(x)] = AB,$

(c) $\lim\limits_{x \to x_0} \dfrac{f(x)}{\varphi(x)} = \dfrac{A}{B}$ *provided that* $\lim\limits_{x \to x_0} \varphi(x) = B \neq 0.$

◄ We shall prove Statement (b) of this theorem.

Let $\lim\limits_{x \to x_0} f(x) = A$ and $\lim\limits_{x \to x_0} \varphi(x) = B.$

By Theorem 7.19 we have

$$f(x) = A + \alpha(x), \qquad \varphi(x) = B + \beta(x),$$

where $\alpha(x)$ and $\beta(x)$ are infinitesimals as $x \to x_0$.

Whence

$$f(x)\varphi(x) = [A + \alpha(x)][B + \beta(x)] = AB + B\alpha(x) + A\beta(x) + \alpha(x)\beta(x).$$

By Theorem 7.15 $B\,\alpha(x)$, $A\,\beta(x)$ and $\alpha(x)\beta(x)$ are infinitesimals as $x \to x_0$. Then their sum is also an infinitesimal as $x \to x_0$. Thus we have represented the function $f(x)\varphi(x)$ as the sum of the constant quantity AB and an infinitesimal as $x \to x_0$. By Theorem 7.19 we conclude that the function $f(x)\varphi(x)$ has a limit equal to AB at x_0, i.e.,

$$\lim_{x \to x_0} [f(x) \cdot \varphi(x)] = AB = \lim_{x \to x_0} f(x) \cdot \lim_{x \to x_0} \varphi(x). \quad ►$$

It is easy to generalize Theorem 7.20 to the case of any finite number of factors multiplied.

Corollary. A constant factor can be taken outside the limit sign.

We leave to the reader to prove Statements (a) and (c) and to solve the following problems.

Problems (1). Let $f(x)$ have a limit at x_0 and let $\varphi(x)$ have no limit at x_0. Which of the following limits exists?

(i) $\lim\limits_{x \to x_0} [f(x) + \varphi(x)]$, (ii) $\lim\limits_{x \to x_0} [f(x) \cdot \varphi(x)]$.

(2) Let $\lim\limits_{x \to x_0} f(x) \neq 0$ and let $\varphi(x)$ have no limit at x_0. Prove that $\lim\limits_{x \to x_0} [f(x) \cdot \varphi(x)]$ does not exist.

Examples. (1) Compute $\lim\limits_{x \to 0} \dfrac{x^2 - 4}{x + 1}$.

◄ Consider the given function as a quotient of two functions $f(x) = x^2 - 4$ and $\varphi(x) = x + 1$ each of which has a limit at $x = 0$, namely, $\lim\limits_{x \to 0} f(x) = \lim\limits_{x \to 0} (x^2 - 4) = -4$ and $\lim\limits_{x \to 0} \varphi(x) = \lim\limits_{x \to 0} (x + 1) = 1 \neq 0$. Since a limit of $\varphi(x)$ is not equal to zero, Statement (c) of Theorem 7.20 is applicable here. Then we have

$$\lim_{x \to 0} \frac{x^2 - 4}{x + 1} = \frac{\lim\limits_{x \to 0} (x^2 - 4)}{\lim\limits_{x \to 0} (x + 1)} = -4. \quad ►$$

(2) Compute $\lim\limits_{x \to 1} \dfrac{x^2 - 1}{x - 1}$.

◄ Setting $f(x) = x^2 - 1$ and $\varphi(x) = x - 1$, we have $\lim\limits_{x \to 1} f(x) = 0$ and $\lim\limits_{x \to 1} \varphi(x) = 0$. Thus we can not use Statement (c) of Theorem 7.20 directly. To resolve this difficulty we can represent the quotient of $f(x)$ and $\varphi(x)$ as

$$\frac{x^2 - 1}{x - 1} = \frac{(x - 1)(x + 1)}{x - 1}.$$

Observing that the definition of a limit at a point does not involve this point and the value of the function at this point, we can exclude the point $x = 1$ from consideration. This means that we can reduce the fraction by $x - 1$ (assuming that $x - 1 \neq 0$). Then we have

$$\frac{x^2 - 1}{x - 1} = x + 1, \; x \neq 1.$$

Whence

$$\lim_{x \to 1} \frac{x^2 - 1}{x - 1} = \lim_{x \to 1} (x + 1) = 2. \;\blacktriangleright$$

(3) Compute $\lim\limits_{x \to 0} \dfrac{\sqrt{1 + x^2} - 1}{x^2}$.

◀ The limits of the nominator and denominator at $x = 0$ are equal to zero, and we can not use Theorem 7.20 directly. But multiplying the nominator and denominator by $\sqrt{1 + x^2} + 1$ and assuming $x \neq 0$, we have at $x \neq 0$

$$\frac{(\sqrt{1 + x^2} - 1)(\sqrt{1 + x^2} + 1)}{x^2(\sqrt{1 + x^2} + 1)} = \frac{x^2}{x^2(\sqrt{1 + x^2} + 1)} = \frac{1}{\sqrt{1 + x^2} + 1}.$$

Then applying the quotient formula given by Statement (c) of Theorem 7.20, we obtain

$$\lim_{x \to 0} \frac{\sqrt{1 + x^2} - 1}{x^2} = \lim_{x \to 0} \frac{1}{\sqrt{1 + x^2} + 1} = \frac{1}{\lim\limits_{x \to 0} (\sqrt{1 + x^2} + 1)} = \frac{1}{2}. \;\blacktriangleright$$

(4) Consider the function $f(x) = \sin x^2$ which is an even function, defined at every point of the number line and bounded everywhere so that $|\sin x^2| \leqslant 1, \; \forall x$. This function vanishes at $x = \pm\sqrt{n\pi}$, where $n = 0, 1, 2, \ldots$.

◀ Consider two consecutive points at which $f(x)$ vanishes. Let these points be $\sqrt{n\pi}$ and $\sqrt{(n + 1)\pi}$ and the distance between them be $\sqrt{(n + 1)\pi} - \sqrt{n\pi}$.

Now we compute $\lim\limits_{n \to \infty} (\sqrt{(n + 1)\pi} - \sqrt{n\pi})$. Multiplying and dividing $(\sqrt{(n + 1)\pi} - \sqrt{n\pi})$ by $(\sqrt{(n + 1)\pi} + \sqrt{n\pi})$, we obtain

$$\lim_{n\to\infty}(\sqrt{(n+1)\pi}-\sqrt{n\pi})$$

$$=\lim_{n\to\infty}\frac{(\sqrt{(n+1)\pi}-\sqrt{n\pi})(\sqrt{(n+1)\pi}+\sqrt{n\pi})}{\sqrt{(n+1)\pi}+\sqrt{n\pi}}$$

$$=\lim_{n\to\infty}\frac{\pi}{\sqrt{(n+1)\pi}+\sqrt{n\pi}}=0,$$

since the nominator of the above fraction is a constant quantity and the denominator increases infinitely as $n \to \infty$. Thus the distance between two consecutive points at which $f(x)$ vanishes tends to zero as $n \to \infty$. Hence, the function $f(x) = \sin x^2$ is nonperiodic. ►

One-sided limits. Let $f(x)$ be defined in an interval (a, x_0). Then a number A is said to be a *left-hand limit* of $f(x)$ at x_0 if, given any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|f(x) - A| < \varepsilon,$$

whenever $x_0 - \delta < x < x_0$.

In symbols, we write

$$A = \lim_{x\to x_0-0} f(x) \quad \text{or} \quad A = f(x_0 - 0).$$

Now let $f(x)$ be defined in an interval (x_0, b). Then a number A is said to be a *right-hand limit* of $f(x)$ at x_0 if, given any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|f(x) - A| < \varepsilon$$

whenever $x_0 < x < x_0 + \delta$.

In symbols, we write

$$A = \lim_{x\to x_0-0} f(x) \quad \text{or} \quad A = f(x_0 + 0).$$

Suppose that a function $f(x)$ is defined in the neighbourhood $(x_0 - h_1, x_0 + h_2)$, $h_1 > 0, h_2 > 0$, of x_0, except probably at x_0. Furthermore, let there exist left- and right-hand limits of $f(x)$ at x_0 (Fig. 7.21).

Then applying the Cauchy criterion for limit of $f(x)$ at x_0, we can easily prove that for the function $f(x)$ to have a limit at x_0 it is necessary and sufficient that there exist left- and right-hand limits of $f(x)$ at x_0 and these limits be equal to each other. In other words, it is easy to prove that

$$f(x_0 - 0) = f(x_0 + 0) = \lim_{x\to x_0} f(x).$$

◄ Let $\lim\limits_{x \to x_0} f(x) = A$. Then given any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|f(x) - A| < \varepsilon \tag{*}$$

holds for all x $(x \neq x_0)$ in $(x_0 - \delta, x_0 + \delta)$.

Evidently, (*) holds for all x in $(x_0 - \delta, x_0)$ and all x in $(x_0, x_0 + \delta)$ so that by definition we have

$$A = \lim_{x \to x_0 - 0} f(x) \text{ and } A = \lim_{x \to x_0 + 0} f(x).$$

Conversely, let $A = \lim\limits_{x \to x_0 - 0} f(x)$ and $A = \lim\limits_{x \to x_0 + 0} f(x)$. Then given any $\varepsilon > 0$ there exist $\delta_1 > 0$ and $\delta_2 > 0$ such that

$$|f(x) - A| < \varepsilon$$

whenever $x_0 - \delta_1 < x < x_0$ and $x_0 < x < x_0 + \delta_2$.

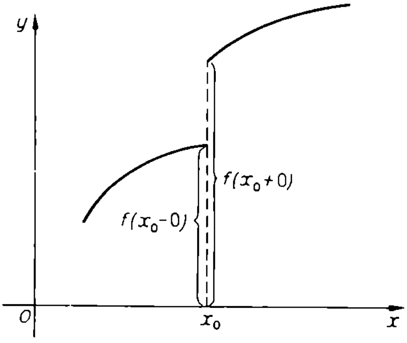


Fig. 7.21

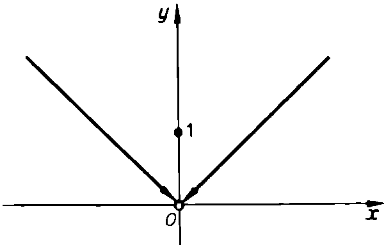


Fig. 7.22

Set $\delta = \min\{\delta_1, \delta_2\}$. Then we obtain

$$|f(x) - A| < \varepsilon$$

for all x such that $0 < |x - x_0| < \delta$. Whence it follows that

$$\lim_{x \to x_0} f(x) = A. \text{ ►}$$

Examples. (1) Consider the function $f(x) = \begin{cases} |x| & \text{for } x \neq 0, \\ 1 & \text{for } x = 0, \end{cases}$ shown in Fig. 7.22.

We have

$$\lim_{x \to 0 - 0} f(x) = \lim_{x \to 0 + 0} f(x) = 0 \Rightarrow \lim_{x \to 0} f(x) = 0.$$

(2) Consider the function $f(x) = \dfrac{1}{1 + e^{1/x}}$, $x \neq 0$, shown in Fig. 7.23.

We have

$$\lim_{x \to 0-0} f(x) = 1 \text{ and } \lim_{x \to 0+0} f(x) = 0 \Rightarrow \bar{\exists} \lim_{x \to 0} f(x).$$

(3) Consider the function $f(x) = e^{1/x}$, $x \neq 0$, shown in Fig. 7.24. We have

$$\lim_{x \to 0-0} f(x) = 0 \text{ and } \lim_{x \to 0+0} f(x) = +\infty \Rightarrow \bar{\exists} \lim_{x \to 0} f(x).$$

In conclusion we suppose that $f(x)$ is defined on a closed interval $[a, b]$ or on an open interval (a, b). Then at a $f(x)$ can have only a right-hand limit and at b it can only have a left-hand limit.

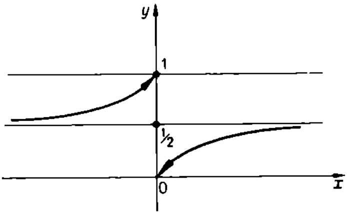


Fig. 7.23

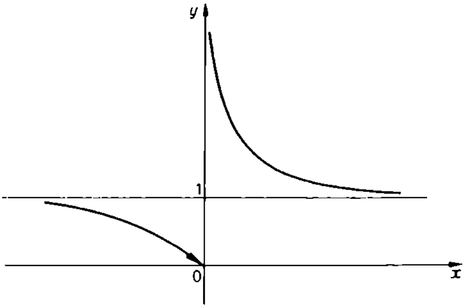


Fig. 7.24

7.6 Continuous Functions. Continuity at a Point

Notion of continuity at a point. Let a function $f(x)$ be defined in a neighbourhood Ω of a point x_0. A function $f(x)$ is continuous at x_0 if (i) $f(x)$ has a limit at x_0 and (ii) the limit of $f(x)$ at x_0 is equal to the value $f(x_0)$ of $f(x)$ at x_0, i.e.,

$$\lim_{x \to x_0} f(x) = f(x_0). \tag{$*$}$$

Since $x_0 = \lim_{x \to x_0} x$ we can write $(*)$ as

$$\lim_{x \to x_0} f(x) = f\left(\lim_{x \to x_0} x\right).$$

In other words, for a continuous function $f(x)$ the symbols lim and f can be interchanged in $(*)$.

Now we give a precise "ε-δ" definition of continuity.

Let $f(x)$ be defined at x_0 and in a neighbourhood Ω of x_0. Then $f(x)$ is a continuous function at x_0 if, given any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|f(x) - f(x_0)| < \varepsilon$$

whenever $|x - x_0| < \delta$ and $x \in \Omega$.

Using logical symbols, we can express this definition as

$(f(x)$ is continuous at $x_0) \Leftrightarrow \forall \varepsilon > 0 \ \exists \delta > 0, \ \forall x \in \Omega,$

$$|x - x_0| < \delta \Rightarrow |f(x) - f(x_0)| < \varepsilon.$$

Notice that in general the value of δ depends on both the number $\varepsilon > 0$ and the point x_0, i.e., $\delta = \delta(\varepsilon, x_0)$. It is important to mention here that the "ε-δ" definition of continuity does not require $x \neq x_0$.

We can also give a slightly different definition of continuity at a point. Let $y = f(x)$ be defined at x_0 and in the neighbourhood Ω of x_0 (Fig. 7.25). Consider the point $x = x_0 + \Delta x$ in Ω which differs from x_0 by a positive or negative quantity Δx. The quantity Δx is called the *increment of the argument* x at x_0. Then the difference

$$\Delta y = f(x_0 + \Delta x) - f(x_0)$$

is called the *increment of the function* f at x_0, corresponding to the increment Δx of the independent variable x.

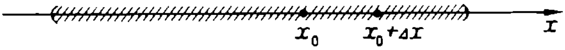


Fig. 7.25

In terms of the increment of x the condition of continuity of $f(x)$ at x_0, i.e.,

$$\lim_{x \to x_0} f(x) = f(x_0)$$

becomes

$$\lim_{\Delta x \to 0} f(x_0 + \Delta x) = f(x_0)$$

or, equivalently,

$$\lim_{\Delta x \to 0} [f(x_0 + \Delta x) - f(x_0)] = 0. \qquad (**)$$

Observe that $f(x_0 + \Delta x) - f(x_0) = \Delta y$. Then $(**)$ can be written as

$$\lim_{\Delta x \to 0} \Delta y = 0,$$

which offers another definition of the continuity of $y = f(x)$ at the point x_0: the function $f(x)$ is continuous at $x_0 \in \Omega$ if the increment of f at x_0, corresponding to the increment Δx of x, tends to zero as $\Delta x \to 0$.

Example. Let us show that the function $y = x^2$ is continuous at every point of the number line.

◀ Indeed, given any increment Δx at x_0 we have

$$\Delta y = (x_0 + \Delta x)^2 - x_0^2 = 2x_0\Delta x + (\Delta x)^2 = (2x_0 + \Delta x)\Delta x.$$

Whence it follows that $\Delta y \to 0$ as $\Delta x \to 0$. This implies that $y = x^2$ is continuous at every point x_0 of the number line. ▶

The definitions given above are equivalent to each other and any of them can be applied when suitable.

In many cases it is convenient to apply the definition of continuity at a point which stems from the sequential criterion for limits. Let $f(x)$ be defined on any set E of real numbers and let $x_0 \in E$. A function $f(x)$ is continuous at x_0 if for any sequence $\{x_n\}$, $x_n \in E$ which converges to x_0 the corresponding sequence $\{f(x_n)\}$ of the values of $f(x)$ converges to $f(x_0)$.

Example. Prove that the Dirichlet function

$$D(x) = \begin{cases} 1 \text{ for rational } x, \\ 0 \text{ for irrational } x \end{cases}$$

is discontinuous at every point.

◀ Let x_0 be any irrational number. Then $D(x_0) = 0$ and whatever x_0 there exists a sequence $\{r_n\}$ of rational numbers, converging to x_0. By definition $D(r_n) = 1 \ \forall n$; so the sequence $\{D(r_n)\} = \{1\}$ converges to 1 so that $\{D(r_n)\} \not\to D(x_0)$. This means that $D(x)$ is not continuous for irrational x. By similar reasoning we can easily verify that $D(x)$ is discontinuous for all rational x. ▶

It is not hard to show that the "ε-δ" definition of continuity at a point and the definition stemming from the sequential criterion for limits are equivalent.

Properties of functions continuous at a point. The following two theorems describe the important properties of continuous functions.

Theorem 7.21. *If a function $f(x)$ is continuous at a point x_0 and if $f(x_0) > A$ ($f(x_0) < A$) then there exists $\delta > 0$ such that $f(x) > A$ ($f(x) < A$) for all x in the open interval $(x_0 - \delta, x_0 + \delta)$.*

◀ For definiteness, we assume that $f(x_0) > A$. Then

$$f(x_0) = A + h,$$

where $h > 0$.

Let $\varepsilon = \frac{h}{2}$. Since $f(x)$ is continuous at x_0 there exists $\delta > 0$ such that for all x

$$|f(x) - f(x_0)| < \frac{h}{2}$$

or

$$-\frac{h}{2} < f(x) - f(x_0) < \frac{h}{2}$$

whenever $|x - x_0| < \delta$. Whence

$$f(x) > f(x_0) - \frac{h}{2} = A + h - \frac{h}{2} = A + \frac{h}{2}.$$

Thus, $f(x) > A$, $\forall x \in (x_0 - \delta, x_0 + \delta)$. ►

From the preceding theorem the following theorem can be derived.

Theorem 7.22. *Let $f(x)$ be continuous at x_0 and let $f(x_0) \neq 0$. Then there exists a neighbourhood $(x_0 - \delta, x_0 + \delta)$ of x_0 such that $f(x)$ does not vanish and is of the same sign at every point of the neighbourhood.*

◄ To prove this theorem it suffices to put $A = 0$ in Theorem 7.21. ►

Continuity of elementary functions. Basic elementary functions are[*]

(a) power function $y = x^\alpha$, where α is an arbitrary real number, $x > 0$;

(b) exponential function $y = a^x$, $a > 0$, $a \neq 1$, $-\infty < x < +\infty$;

(c) logarithmic function $y = \log_a x$, $a > 0$, $a \neq 1$, $x > 0$,

(d) trigonometric functions $y = \sin x$, $-\infty < x < +\infty$, $y = \cos x$; $-\infty < x < +\infty$, $y = \tan x$, $x \neq \frac{\pi}{2} + n\pi$, $n = 0, \pm 1, \pm 2, \ldots$, $y = \cot x$; $x \neq n\pi$, $n = 0, \pm 1, \pm 2, \ldots$;

(e) inverse trigonometric functions $y = \sin^{-1} x$, $-1 \leqslant x \leqslant 1$, $y = \cos^{-1} x$, $-1 \leqslant x \leqslant 1$, $y = \tan^{-1} x$, $-\infty < x < +\infty$, $y = \cot^{-1} x$, $-\infty < x < +\infty$.

We say that *elementary functions* are those obtained by means of a finite number of arithmetic operations and a finite number of operations of constructing a function of a function. It is not hard to show that the basic elementary functions are continuous at every point of their domains.

We prove that the function $y = \cos x$ is continuous at every point of the domain $-\infty < x < +\infty$.

◄ First we prove the auxiliary inequality

$$|\sin x| < |x| \ \forall x. \qquad (*)$$

Look at a circle of unit radius (Fig. 7.26). Let a radian measure of the angle AOB be x, $0 < x < \frac{\pi}{2}$ and let the angle AOB be equal to the angle AOC. Evidently, the length of the segment BC is equal to $2 \sin x$ and the length of the arc $\breve{BC}$ is equal to $2x$. Since the arc length is greater than the length of the subtending chord we have

---

[*] A detailed account of basic elementary functions is given in the Appendix.

$$2 \sin x < 2x \text{ and } \sin x < x.$$

This inequality can be written as $|\sin x| < |x|$ for $x \in (0, \pi/2)$. The identities $|\sin(-x)| = |-\sin x| = |\sin x|$ and $|-x| = |x|$ imply that the inequality $|\sin x| < |x|$ also holds for $x \in (-\pi/2, 0)$. Observe that (*) is also true for all $x \in (-\pi/2, \pi/2)$ since $\sin 0 = 0$. If $x \notin (-\pi/2, \pi/2)$ then $|x| \geqslant \pi/2 > 1$, whereas $|\sin x| \leqslant 1$ $\forall x$. Thus we conclude that the inequality

$$|\sin x| \leqslant |x|$$

holds for any x.

Now we turn to the function $y = \cos x$ which is defined at each point of the number line. Let x be an arbitrary point ($x \in \mathbb{R}$) and Δx be an increment of x at x. Then the increment of $y = \cos x$ can be expressed as

$$\Delta y = \cos(x + \Delta x) - \cos x = -2 \sin\left(x + \frac{\Delta x}{2}\right) \sin\frac{\Delta x}{2}.$$

Whence

$$\begin{aligned} |\Delta y| &= \left| -2 \sin\left(x + \frac{\Delta x}{2}\right) \sin\frac{\Delta x}{2} \right| \\ &= 2 \left| \sin\left(x + \frac{\Delta x}{2}\right) \right| \left| \sin\frac{\Delta x}{2} \right|. \qquad (**) \end{aligned}$$

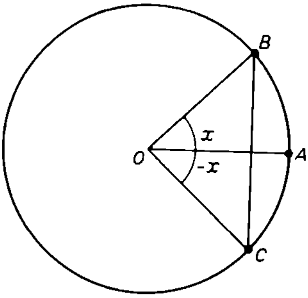


Fig. 7.26

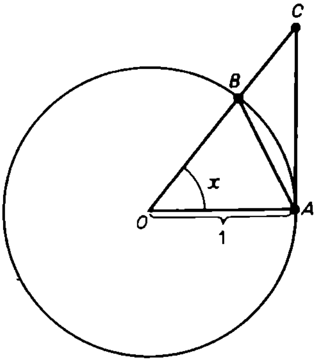


Fig. 7.27

Recall that $\left| \sin\left(x + \frac{\Delta x}{2}\right) \right| \leqslant 1$ $\forall x$, $\forall \Delta x$ and, by virtue of (*), $\left| \sin\frac{\Delta x}{2} \right| \leqslant \frac{|\Delta x|}{2}$. Thus we can write (**) as

$$|\Delta y| \leqslant 2 \cdot 1 \cdot \frac{|\Delta x|}{2} = |\Delta x|$$

and

$$0 \leqslant |\Delta y| \leqslant |\Delta x|,$$

where Δy is a function of Δx for a given x.

Whence, by virtue of Theorem 7.13, we obtain

$$\lim_{\Delta x \to 0} \Delta y = 0$$

which means that the function $y = \cos x$ is continuous at every point of the number line. ►

Remarkable limits. We consider the relation $\dfrac{\sin x}{x}$ which is a function of x defined at any $x \neq 0$ and prove the following theorem.

Theorem 7.23. *If* x *is measured in radians then*

$$\lim_{x \to 0} \frac{\sin x}{x} = 1.$$

◄ Suppose that $0 < x < \pi/2$ and consider a circle of unit radius (Fig. 7.27). From Fig. 7.27 it is easily seen that

$$\text{area of } \Delta OAB < \text{ area of sector } OAB < \text{ area of } \Delta OAC.$$

These areas are equal to $\frac{1}{2}\sin x$, $\frac{1}{2}x$ and $\frac{1}{2}\tan x$, respectively. Then we have

$$\sin x < x < \tan x, \quad x \in (0, \ \pi/2).$$

Dividing all the terms of these inequalities by $\sin x > 0$, we obtain

$$1 < \frac{x}{\sin x} < \frac{1}{\cos x}.$$

Whence

$$1 > \frac{\sin x}{x} > \cos x \tag{***}$$

which holds for $x \in (0, \ \pi/2)$ and also for $x \in (-\pi/2, \ 0)$ since

$$\frac{\sin(-x)}{-x} = \frac{\sin x}{x} \text{ and } \cos(-x) = \cos x.$$

The function $y = \cos x$ is continuous at any x and, in particular, at $x = 0$ so that

$$\lim_{x \to 0} \cos x = \cos 0 = 1. \tag{****}$$

Thus both $\varphi(x) \equiv \cos x$ and $\psi(x) \equiv 1$ have the limits equal to 1 at $x = 0$. By Theorem 7.13 (***) and (****) yield the identity we are seeking for. ►

In the preceding sections we have proved that

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = \mathrm{e}.$$

Denoting $1/n = z$ we have $n = 1/z$ and $z \to 0$ as $n \to \infty$. Then

$$\lim_{z\to 0}(1+z)^{1/z} = \mathrm{e},$$

where $z = 1, 1/2, 1/3, \ldots$.

We can verify that there exists $\lim\limits_{x\to 0}(1+x)^{1/x}$ which is also equal to the number e when x tends to 0 in an arbitrary way by running through any sequence of real values distinct from 0, i.e.,

$$\lim_{x\to 0}(1+x)^{1/x} = \mathrm{e}.$$

Operations on functions continuous at a point. In this section we shall develop a number of important properties of continuous functions.

Theorem 7.24. *Let $f(x)$ and $\varphi(x)$ be defined at x_0 and in a neighbourhood of x_0. If $f(x)$ and $\varphi(x)$ are continuous at x_0 so are the sum $f(x) + \varphi(x)$, the difference $f(x) - \varphi(x)$, the product $f(x)\varphi(x)$ and, if $\varphi(x) \neq 0$, the quotient $\dfrac{f(x)}{\varphi(x)}$.*

By way of illustration we shall prove the quotient rule; all the other rules can be proved in a similar way.

◄ Let $f(x)$ and $\varphi(x)$ be continuous at x_0 and let $\varphi(x_0) \neq 0$. By Theorem 7.22 there exists a neighbourhood of x_0 such that $\varphi(x) \neq 0$ for all x in this neighbourhood. Thus, the function $F(x) = \dfrac{f(x)}{\varphi(x)}$ is defined at x_0 and in a neighbourhood of x_0. Since $\lim\limits_{x\to x_0} f(x) = f(x_0)$ and $\lim\limits_{x\to x_0}\varphi(x) = \varphi(x_0) \neq 0$ the quotient rule (see Theorem 7.20) is applicable here and we have

$$\lim_{x\to x_0} F(x) = \lim_{x\to x_0}\frac{f(x)}{\varphi(x)} = \frac{\lim\limits_{x\to x_0} f(x)}{\lim\limits_{x\to x_0}\varphi(x)} = \frac{f(x_0)}{\varphi(x_0)} = F(x_0).$$

Thus $\lim\limits_{x\to x_0} F(x) = F(x_0)$, i.e., the function $F(x) = \dfrac{f(x)}{\varphi(x)}$ is continuous at x_0. ►

Composite functions. Let E be an arbitrary subset of a number line and let $u = \varphi(x)$ be a function defined on E. We denote by E_1 a set of values of u corresponding to the values of x in E. Furthermore, let there be defined a function $y = f(u)$ on E_1. Then every $x \in E$ corresponds to a certain $u \in E_1$ which in turn is associated with a certain value of $y = f(u)$ (Fig. 7.28). Thus, the value of y is a function of x which is defined on E. In this case we say that y is a composite function of x and write

$$y = f[\varphi(x)].$$

For instance, if $u = \sin x$ and $y = e^u$ then $y = e^{\sin x}$ is a composite function defined at all x.

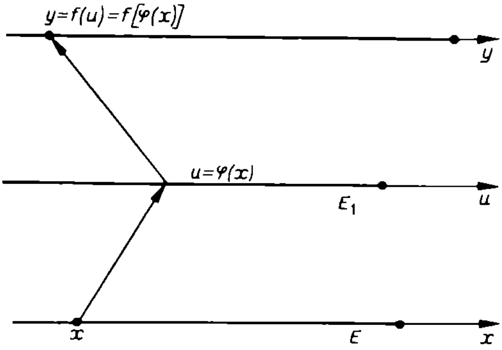


Fig. 7.28

Theorem 7.25. *Let $u = \varphi(x)$ have a limit equal to A at x_0 and let $y = f(u)$ be a continuous function at the point $u = A$. Then a composite function $y = f\,[\varphi(x)]$ has a limit equal to $f(A)$ at x_0.*

◄ We have to prove that $\lim\limits_{x \to x_0} f[\varphi(x)] = f(A)$, i.e.,

$$\forall \varepsilon > 0 \ \exists \delta > 0, \ \forall x, \ x \neq x_0, \ |x - x_0| < \delta \Rightarrow |f[\varphi(x)] - f(A)| < \varepsilon.$$

Let $\varepsilon > 0$ be an arbitrary number. Since $f(u)$ is continuous at $u = A$ then for the given $\varepsilon > 0$ there exists a number $\eta > 0$ such that

$$|f(u) - f(A)| < \varepsilon \qquad (*)$$

for all u whenever $|u - A| < \eta$.

By definition $A = \lim\limits_{x \to x_0} \varphi(x)$ so that for any $\eta > 0$ there exists $\delta > 0$ such that $|\varphi(x) - A| < \eta$ or, equivalently, $|u - A| < \eta$ for all $x \neq x_0$ whenever $|x - x_0| < \delta$.

The inequality $|u - A| < \eta$ implies $(*)$ which can be written as

$$|f[\varphi(x)] - f(A)| < \varepsilon.$$

Thus, for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|f[\varphi(x)] - f(A)| < \varepsilon$$

whenever $|x - x_0| < \delta$ and $x \neq x_0$.

This means that the number $f(A)$ is a limit of the composite function $f[\varphi(x)]$ at x_0. ►

We see that when the conditions of Theorem 7.25 are fulfilled we have

$$\lim_{x\to x_0} f[\varphi(x)] = f(A)$$

or, equivalently,

$$\lim_{x\to x_0} f[\varphi(x)] = f[\lim_{x\to x_0} \varphi(x)].$$

The latter identity indicates the rule of computing a limit of a composite function.

Example. Show that $\lim\limits_{x\to 0}\dfrac{\ln(1+x)}{x} = 1.$

◄ Observe that $\dfrac{\ln(1+x)}{x} = \ln(1+x)^{1/x}$. The function $y = \ln(1+x)^{1/x}$ is a composite function made up of the functions $y = \ln u$ and $u = (1+x)^{1/x}$. Since $\lim\limits_{x\to 0}(1+x)^{1/x} = \mathrm{e}$ and $y = \ln u$ is continuous at $u = \mathrm{e}$. Theorem 7.25 gives

$$\lim_{x\to 0}\frac{\ln(1+x)}{x} = \lim_{x\to 0}\ln(1+x)^{1/x} = \ln[\lim_{x\to 0}(1+x)^{1/x}] = \ln \mathrm{e} = 1. \quad ►$$

Theorem 7.26. *Let $u = \varphi(x)$ be a continuous function at x_0 and let $y = f(u)$ be continuous at $u_0 = \varphi(x_0)$. Then a composite function $y = f[\varphi(x)]$ is continuous at x_0.*

◄ Since $u = \varphi(x)$ is continuous at x_0 it has at x_0 a limit equal to $\varphi(x_0) = u_0$. Besides, $y = f(u)$ is continuous at u_0. Then Theorem 7.25 implies that the composite function $y = f[\varphi(x)]$ has at x_0 a limit equal to $f(u_0) = f[\varphi(x_0)]$ or

$$\lim_{x\to x_0} f[\varphi(x)] = f[\varphi(x_0)].$$

Whence we infer that a composite function $f[\varphi(x)]$ is continuous at x_0. ►

Discontinuities of functions. Let $f(x)$ be defined at a point x_0 and in a neighbourhood of x_0. If $f(x)$ is continuous at x_0 then

$$\lim_{x\to x_0} f(x) = f(x_0)$$

or, in terms of one-sided limits,

$$\lim_{x\to x_0-0} f(x) = \lim_{x\to x_0+0} f(x) = f(x_0). \tag{*}$$

Thus the function $f(x)$ is continuous at the point x_0 if and only if there exist the left-hand and right-hand limits of $f(x)$ at x_0 these limits being equal to each other and to the value of $f(x)$ at x_0.

Definition. The function $f(x)$ has a discontinuity at the point x_0 if $f(x)$ is not continuous at x_0 and x_0 is called the *point of discontinuity*[*].

The discontinuities of a function can be classified according to the sense of violating the condition (*).

Definition. A point x_0 is called a *removable discontinuity* of $f(x)$ if $f(x)$ has the left- and right-hand limits at x_0 which are equal to each other but are different from the value of $f(x)$ at x_0, i.e.,

$$\lim_{x\to x_0-0} f(x) = \lim_{x\to x_0+0} f(x) \neq f(x_0).$$

When x_0 is a removable discontinuity it suffices to modify the function $f(x)$ only at the point x_0 so that $f(x)$ becomes continuous at x_0. That is, if x_0 is a removable discontinuity of $f(x)$ then the function

$$F(x) = \begin{cases} f(x) & \text{for } x \neq x_0, \\ \lim\limits_{x\to x_0} f(x) & \text{for } x = x_0 \end{cases}$$

is continuous at x_0. We have removed a discontinuity by changing the value of $f(x)$ at one point, x_0.

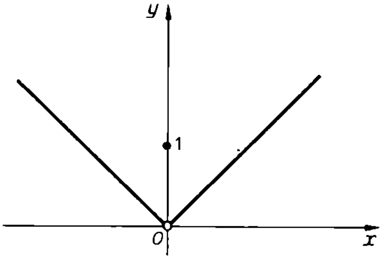


Fig. 7.29

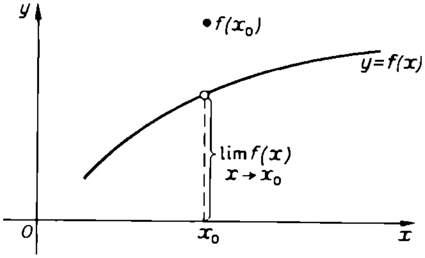


Fig. 7.30

Example. Let

$$f(x) = \begin{cases} |x| & \text{for } x \neq 0, \\ 1 & \text{for } x = 0. \end{cases}$$

◄ We have

$$\lim_{x\to x_0-0} f(x) = \lim_{x\to x_0+0} f(x) = 0 \neq 1 = f(0),$$

so that the point $x = 0$ is a removable discontinuity of $f(x)$ (Fig. 7.29). If we modify $f(x)$ at $x = 0$ by putting $f(0) = 0$ then we shall get a function $F(x) = |x|$ which is continuous at $x = 0$. ►

---

[*] If $f(x)$ is not defined at x_0 the point x_0 is also called the point of discontinuity.

In general, if a function $f(x)$ is continuous on an open interval $(x_0 - \delta_1, x_0 + \delta_2)$ everywhere, except at x_0, and if x_0 is a removable discontinuity then the graph of $f(x)$ is represented by a continuous curve from which the point corresponding to the abscissa x_0 is deleted and is replaced by the point of $f(x_0)$ whose ordinate is different from the ordinate of the removed point (Fig. 7.30). It is important to mention that if x_0 is a removable discontinuity then at x_0 there exists $\lim\limits_{x \to x_0} f(x)$.

We say that x_0 is an *unremovable discontinuity* of $f(x)$ if $\lim\limits_{x \to x_0} f(x)$ does not exist at x_0.

Definition. If the left- and right-hand limits of $f(x)$ at x_0 are finite and are different, i.e., if

$$\lim_{x \to x_0 - 0} f(x) \neq \lim_{x \to x_0 + 0} f(x),$$

then the point x_0 is a discontinuity at which $f(x)$ has a *jump* (or a *gap*) and $f(x_0)$ does not need to be equal to either of the one-sided limits. In this case the jump of $f(x)$ at x_0 is equal to the difference $f(x_0 + 0) - f(x_0 - 0)$ of the left- and right-hand limits of $f(x)$ at x_0.

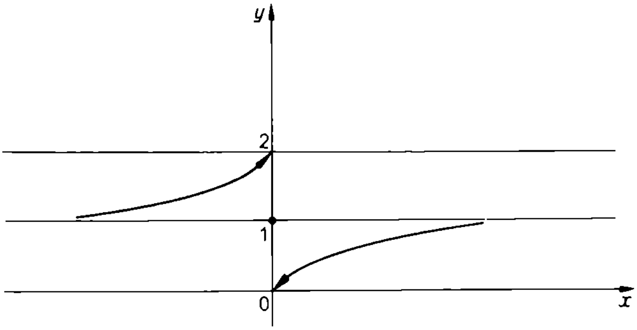


Fig. 7.31

Example. Consider $f(x) = \dfrac{2}{1 + e^{1/x}}$, $f(0) = 1$ (Fig. 7.31).

◄ At $x = 0$ the function $f(x)$ has a jump since

$$\lim_{x \to 0 - 0} f(x) = 2 \quad \text{and} \quad \lim_{x \to 0 + 0} f(x) = 0. \ ►$$

A removable discontinuity and a discontinuity at which the function has a jump are called the *discontinuities of the first kind.* If x_0 is a discontinuity of the first kind the function $f(x)$ has finite left-hand and right-hand limits at x_0. All other discontinuities are called *discontinuities of the second* *kind.* Thus, if x_0 is a discontinuity of the second kind the function $f(x)$ at x_0 does not have either a finite left-hand limit or a finite right-hand limit.

Examples. (1) If $f(x) = \frac{1}{x}$, $x \neq 0$, $f(0) = 0$, the point $x = 0$ is a discontinuity of the second kind, for $\lim\limits_{x \to 0-0} f(x) = -\infty$ and $\lim\limits_{x \to 0+0} f(x) = +\infty$.

(2) Let $f(x) = \sin \frac{1}{x}$, $x \neq 0$, $f(0) = 0$.

◄ At $x = 0$ $f(x)$ has neither finite nor infinite one-sided limits. We can easily prove this by applying the sequential criterion for limits. Thus the point $x = 0$ is a discontinuity of the second kind. ►

(3) For the Dirichlet function

$$D(x) = \begin{cases} 1 & \text{for rational } x, \\ 0 & \text{for irrational } x \end{cases}$$

any point x_0 is a discontinuity of the second kind.

We say that a function $f(x)$ is *continuous at a point* x_0 *from the right* if

$$\lim_{x \to x_0+0} f(x) = f(x_0) \quad \text{or} \quad f(x_0 + 0) = f(x_0),$$

and $f(x)$ is *continuous at* x_0 *from the left* if

$$\lim_{x \to x_0-0} f(x) = f(x_0) \quad \text{or} \quad f(x_0 - 0) = f(x_0).$$

7.7 Continuity on a Closed Interval

Definitions. A function $f(x)$ is said to be *continuous on an open interval* (a, b) if $f(x)$ is continuous at every point of (a, b). We denote a set of all functions continuous on an open interval (a, b) by $C(a, b)$.

A function $f(x)$ is said to be *continuous on a closed interval* $[a, b]$ if $f(x)$ is continuous on an open interval (a, b) and is continuous from the left at b and from the right at a. We denote a set of all functions continuous on $[a, b]$ by $C[a, b]$.

Properties of functions continuous on closed intervals can be established with the help of the following theorems.

Theorem 7.27. *Let* $f(x)$ *be continuous on a closed interval* $[a, b]$ *and let* $f(x)$ *and* $f(b)$ *be of opposite signs. Then there exists at least one point in* (a, b) *at which* $f(x)$ *vanishes.*

◄ Let $f(a)$ and $f(b)$ be of opposite signs. The point $\xi = \frac{a + b}{2}$ is a midpoint of $[a, b]$. If $f(\xi) = 0$ we get what we are seeking for, i.e., the theorem is true. Let $f(\xi) \neq 0$. Consider the closed intervals $[a, \xi]$ and $[\xi, b]$. One of these intervals denoted by $[a_1, b_1]$ is such that $f(x)$ assumes the values of opposite signs at its endpoints. Clearly, the point $\xi_1 = \dfrac{a_1 + b_1}{2}$ becomes the midpoint of $[a_1, b_1]$. If $f(\xi_1) = 0$ the theorem is proved. However it may occur that $f(\xi_1) \neq 0$. Then we choose between the closed intervals $[a_1, \xi_1]$ and $[\xi_1, b_1]$ the one at whose endpoints $f(x)$ assumes the values of opposite signs, denote this closed interval by $[a_2, b_2]$, and repeat the process as above.

It is clear that if we continue this process we finally either arrive at a point $\alpha \in (a, b)$ where $f(\alpha) = 0$, thus completing the proof of the theorem, or we obtain an infinite sequence of the nested closed intervals

$$[a, b] \supset [a_1, b_1] \supset \dots \supset [a_n, b_n] \supset \dots .$$

All the closed intervals tend to zero as $n \to \infty$, i.e.,

$$\lim_{x \to \infty} (b_n - a_n) = \lim_{n \to \infty} \frac{b - a}{2^n} = 0$$

and on every closed interval $f(x)$ assumes the values of opposite signs at the endpoints.

Recall that the Cantor lemma implies that there exists a unique point α contained in each $[a_n, b_n]$, $n = 0, 1, 2, \dots$. Let us prove that $f(\alpha) = 0$.

We assume the converse, i.e., $f(\alpha) \neq 0$, and for definiteness we set $f(\alpha) > 0$. Since $f(x)$ is continuous at $\alpha \in [a, b]$. Theorem 7.22 implies that there exists an open interval $(\alpha - \delta, \alpha + \delta)$ in which $f(x)$ is positive. As $\alpha = \lim\limits_{n \to \infty} a_n = \lim\limits_{n \to \infty} b_n$ we can choose sufficiently large n so that the closed interval $[a_n, b_n]$ is contained in the open interval $(\alpha - \delta, \alpha + \delta)$. This means that $f(a_n)$ and $f(b_n)$ are of the same sign (both positive). However on every closed interval $[a_n, b_n]$ $f(a_n)$ and $f(b_n)$ are of opposite signs. Thus we have a contradiction which makes our assumption of $f(\alpha) \neq 0$ false. Hence, $f(\alpha) = 0$. Notice that $a < \alpha < b$, $\alpha \in [a, b]$ and the point α must be distinct from both a and b since $f(a) \neq 0$ and $f(b) \neq 0$. ►

It is easy to interpret this theorem geometrically (Fig. 7.32). If $f(a)f(b) < 0$ then the points $A(a, f(a))$ and $B(b, f(b))$ occur on different semiplanes relative to the x-axis, and the graph of the continuous function $f(x)$ intersects the x-axis at least at one point between a and b.

It is important that $f(x)$ must be continuous of $[a, b]$. If the function is discontinuous at a point in $[a, b]$ it may have a jump from a negative value to a positive one without assuming a zero value as does the function

$$f(x) = \begin{cases} -1 & \text{for} \quad -1 \leqslant x < 0, \\ 1 & \text{for} \quad 0 \leqslant x \leqslant 1, \end{cases}$$

shown in Fig. 7.33.

To illustrate how Theorem 7.27 works we consider the polynomial equation of odd degree with real coefficients

$$P_{2n+1}(x) \equiv a_0x^{2n+1} + a_1x^{2n} + \dots + a_{2n+1} = 0.$$

◄ Let $a_0 > 0$. For sufficiently large absolute values of x $P_{2n+1}(x)$ becomes negative when x is negative and positive when x is positive. Since the polynomial is everywhere a continuous function of x, $P_{2n+1}(x)$ necessarily vanishes at some intermediate point of its domain when changing its sign from negative to positive. Hence any polynomial of an odd degree with real coefficients has at least one real root. ►

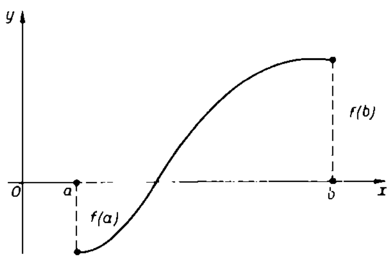


Fig. 7.32

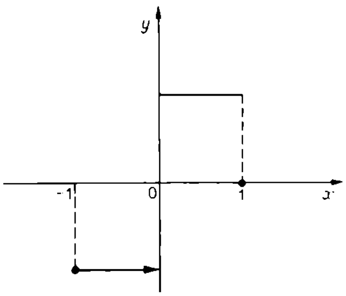


Fig. 7.33

We can apply Theorem 7.27 when we wish to find out whether a polynomial has a real root and, if any, to compute an approximate value of the root.

Example. Consider the polynomial $P_3(x) = x^3 + x - 1$.

◄ Since $P_3(x)$ is of odd degree it has at least one real root.

Evidently, we have $P_3(0) = -1 < 0$ and $P_3(1) = 1 > 0$, i.e., at the endpoints of $[0, 1]$ $P_3(x)$ assumes the values of opposite signs. Hence $P_3(x)$ has a real root contained in $(0, 1)$.

At the midpoint $\xi_1 = 1/2$ of the closed interval $[0, 1]$ we have $P_3(1/2) = -3/8 < 0$. Recall that $P_3(1) > 0$. This means that the desired root is contained in the open interval $(1/2, 1)$. Again we compute the value of $P_3(x)$ at the midpoint of $[1/2, 1]$, i.e., at $\xi_2 = 3/4$. We obtain $P_3(1/2) < 0$ and $P_3(3/4) = 11/64 > 0$. Hence the root of $P_3(x)$ is contained in $(1/2, 3/4)$. Working through this process we obtain a sequence of the nested open intervals of decreasing lengths. In other words, at each step of this process we improve the approximation of the root whose precision is given by the length of the last open interval thus obtained. ►

Theorem 7.28 (intermediate value theorem). *Let $f(x)$ be continuous on $[a, b]$ and let $f(a) = A$ and $f(b) = B$. If C is any value between A and B then there exists at least one point $\alpha \in [a, b]$ such that $f(\alpha) = C$, i.e., if $f(x)$ is continuous on $[a, b]$ it assumes all intermediate values between $f(a)$ and $f(b)$.*

◄ Consider the function $\varphi(x) = f(x) - C$.

For definiteness we assume that $A < B$ and $A < C < B$. Clearly, $\varphi(x)$ is continuous on $[a, b]$ and

$$\varphi(a) = f(a) - C = A - C < 0,$$

$$\varphi(b) = f(b) - C = B - C > 0.$$

By virtue of Theorem 7.27 there exists $\alpha \in (a, b)$ such that $\varphi(\alpha) = f(\alpha) - C = 0$. Hence $f(\alpha) = C$. ►

Theorem 7.29. *If a function $f(x)$ is continuous on $[a, b]$ then $f(x)$ is bounded on $[a, b]$, i.e., there exists a number $K > 0$ such that for all $x \in [a, b]$*

$$|f(x)| \leqslant K.$$

Remark. If a function $f(x)$ is continuous on an open interval (a, b) or on half-open intervals $(a, b]$ or $[a, b)$ then $f(x)$ is not necessarily bounded on the respective interval. For example, $f(x) = 1/x$ is continuous on the half-open interval $(0, 1]$ and is not bounded on it.

Now we suppose that $f(x)$ is defined and bounded on a set E. We say that a supremum of the set of values of $f(x)$ on E is a *supremum M of the function $f(x)$* on E, i.e.,

$$M = \sup_{x \in E} f(x).$$

Similarly, an *infimum m of $f(x)$* on E is

$$m = \inf_{x \in E} f(x).$$

Theorem 7.30. *If a function $f(x)$ is continuous on $[a, b]$ then $f(x)$ attains its supremum and its infimum on $[a, b]$, i.e., there exist numbers ξ and η in $[a, b]$ such that*

$$f(\xi) = m = \inf_{x \in [a,b]} f(x), \quad f(\eta) = M = \sup_{x \in [a,b]} f(x),$$

and

$$f(\xi) \leqslant f(x) \leqslant f(\eta).$$

It is important here that $f(x)$ is continuous on $[a, b]$. For instance, $f(x) = x$ is continuous and bounded on $(-1, 1)$ but does not attain its supremum $\sup\limits_{x\in(-1,1)} x = 1$, i.e., there exists no $x_0 \in (-1, 1)$ such that $f(x_0) = 1$. Analogously, $f(x)$ does not attain its infimum $\inf\limits_{x\in(-1,1)} x = -1$.

Example. Consider $f(x) = x - [x]$ on $[0, 1]$ (Fig. 7.34).

◄ Here $\sup\limits_{x\in[0,1]} f(x) = 1$ is unattainable on $[0, 1]$, since $f(x)$ is not continuous on $[0, 1]$. ►

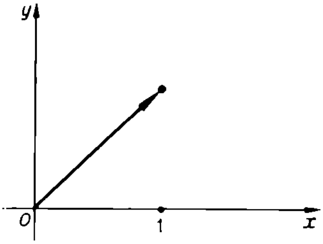


Fig. 7.34

We call the supremum of $f(x)$ the ***absolute maximum*** of $f(x)$ on $[a, b]$ and the infimum of $f(x)$ the ***absolute minimum*** of $f(x)$ on $[a, b]$. Then we can express Theorem 7.30 in a different way.

Theorem 7.30′. *If a function $f(x)$ is continuous on $[a, b]$ then $f(x)$ attains its absolute maximum and absolute minimum on $[a, b]$.*

Uniform continuity. Functions continuous on closed intervals possess an important property of *uniform continuity.*

Let $f(x)$ be continuous on (a, b). Then given any $x_0 \in (a, b)$ and any number $\varepsilon > 0$ there exists a number $\delta > 0$ such that $|f(x) - f(x_0)| < \varepsilon$ whenever $x \in (a, b)$ and $|x - x_0| < \delta$.

The value of δ may depend both on ε and on x_0, i.e., $\delta = \delta(\varepsilon, x_0)$ so that it may happen that for a given $\varepsilon > 0$ δ can be different for different $x_k \in (a, b)$ and there exists no δ which will work for all $x_k \in (a, b)$. The requirement that there exist $\delta = \delta(\varepsilon) > 0$ for all x_k is stronger than the requirement that $f(x)$ be continuous on (a, b).

Definition. A function $f(x)$ is said to be *uniformly continuous* on an open interval (a, b) if for every $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that

$$|f(x') - f(x'')| < \varepsilon$$

whenever $|x' - x''| < \delta$ and $x', x'' \in (a, b)$.

Using logical symbols, we can write

$$(f(x) \text{ is uniformly continuous on } (a, b)) \Leftrightarrow \forall\varepsilon > 0,$$
$$\exists\delta = \delta(\varepsilon) > 0,\ \forall x', x'' \in (a, b),$$
$$|x' - x''| < \delta \Rightarrow |f(x') - f(x'')| < \varepsilon.$$

It is important here that for every $\varepsilon > 0$ there exists $\delta > 0$ such that $|f(x') - f(x'')| < \varepsilon$ for all x', x'' belonging to (a, b) only under the condition $|x' - x''| < \delta$. For example, the function $f(x) = x$ is uniformly continuous on the number line. To verify this it suffices to put $\delta = \varepsilon$.

It is easy to see that if $f(x)$ is a uniformly continuous function on (a, b), it is continuous at every point $x \in (a, b)$. The converse is not true. For instance, $f(x) = \sin \frac{\pi}{x}$ is continuous on $(0, 1)$ and is not uniformly continuous on $(0, 1)$. Indeed, let $x_n' = \frac{1}{n}$ and $x_n'' = \frac{2}{2n+1}$. Then we can choose n so that the value of

$$|x_n' - x_n''| = \left| \frac{1}{n} - \frac{2}{2n+1} \right| = \frac{1}{n(2n+1)}$$

becomes smaller than any $\delta > 0$ whereas

$$|f(x_n') - f(x_n'')| = 1 > \varepsilon \quad \forall \varepsilon < 1.$$

Thus for $\varepsilon > 0$, say for $\varepsilon = \frac{1}{2}$, and for any $\delta > 0$ there exist points x_n' and x_n'' in $(0, 1)$ such that $|x_n' - x_n''| < \delta$ and $|f(x_n') - f(x_n'')| > \varepsilon$. Hence $f(x) = \sin \frac{\pi}{x}$ is not uniformly continuous on $(0, 1)$.

Example. The function $f(x) = \frac{1}{x}$ is continuous on $(0, 1)$ but is not uniformly continuous on $(0, 1)$.

◄ Let $x_n' = \frac{1}{n}$ and $x_n'' = \frac{1}{n + 2\varepsilon}$, where $\varepsilon > 0$ is an arbitrary number. Then $|x_n' - x_n''| = \frac{2\varepsilon}{n(n+2\varepsilon)}$ which can be made smaller than any $\delta > 0$ by choosing sufficiently large n. However $|f(x_n') - f(x_n'')| = 2\varepsilon > \varepsilon \ \forall \varepsilon > 0$. Hence $f(x) = \frac{1}{x}$ is not uniformly continuous on $(0, 1)$. ▸

The following theorem is worth mentioning.

Theorem 7.31 (due to Cantor). *If a function $f(x)$ is continuous on a closed interval $[a, b]$ then $f(x)$ is uniformly continuous on $[a, b]$.*

7.8 Comparison of Infinitesimals

Definitions. Let $\alpha(x)$ and $\beta(x)$ be infinitesimals as $x \to x_0$. Then (a) $\alpha(x)$ is an infinitesimal of *higher order* than $\beta(x)$ if

$$\lim_{x \to x_0} \frac{\alpha(x)}{\beta(x)} = 0.$$

In symbols, we write $\alpha(x) = o(\beta(x))$, $x \to x_0$, to mean that $o(\beta(x))$, $x \to x_0$, is an infinitesimal of higher order at x_0 than an infinitesimal $\beta(x)$ at x_0.

For example, $\alpha(x) = x^2$ and $\beta(x) = x$ are infinitesimals as $x \to 0$. Then

$$\lim_{x\to 0} \frac{\alpha(x)}{\beta(x)} = \lim_{x\to 0} \frac{x^2}{x} = \lim_{x\to 0} x = 0.$$

Whence $\alpha(x) = o(\beta(x))$, $x \to 0$, or $x^2 = o(x)$, $x \to 0$.

(b) $\alpha(x)$ and $\beta(x)$ are infinitesimals of the *same order* if

$$\lim_{x\to x_0} \frac{\alpha(x)}{\beta(x)} = C \neq 0.$$

For example, $\alpha(x) = 2x$ and $\beta(x) = x$, $x \to 0$ are infinitesimals of the same order since

$$\lim_{x\to 0} \frac{\alpha(x)}{\beta(x)} = \lim_{x\to 0} \frac{2x}{x} = 2.$$

(c) $\alpha(x)$ and $\beta(x)$ are *incomparable infinitesimals* as $x \to x_0$ if the ratio $\dfrac{\alpha(x)}{\beta(x)}$ has neither finite nor infinite limit as $x \to x_0$. For example, $\alpha(x) = x \sin \dfrac{1}{x}$ and $\beta(x) = x$ are incomparable since the ratio $\dfrac{\alpha(x)}{\beta(x)} = \sin \dfrac{1}{x}$ has no finite limit at $x = 0$ and is not an infinity as $x \to 0$.

(d) $\alpha(x)$ is an infinitesimal of the mth order, where m is a positive integer, relative to the infinitesimal $\beta(x) = x - x_0$ as $x \to x_0$ if

$$\lim_{x\to x_0} \frac{\alpha(x)}{(x - x_0)^m} = B \neq 0.$$

For example, $\alpha(x) = 3\sin^2 x$ is an infinitesimal of order $m = 2$ relative to $\beta(x) = x$ as $x \to 0$ since

$$\lim_{x\to 0} \frac{3\sin^2 x}{x^2} = 3.$$

Equivalent infinitesimals. The infinitesimals $\alpha(x)$ and $\beta(x)$ as $x \to x_0$ are said to be *equivalent* if at the point x_0

$$\lim_{x\to x_0} \frac{\alpha(x)}{\beta(x)} = 1.$$

Thus equivalent infinitesimals are infinitesimals of the same order.

We write

$$\alpha(x) \sim \beta(x),\ x \to x_0$$

to mean that $\alpha(x)$ and $\beta(x)$ are equivalent infinitesimals.

Sometimes equivalent infinitesimals $\alpha(x)$ and $\beta(x)$ are called *asymptotically equal* as $x \to x_0$.

Remark. Let $\alpha(x)$, $\beta(x)$ and $\gamma(x)$ be infinitesimals as $x \to x_0$. It is not hard to verify that

(1) $\alpha(x) \sim \alpha(x)$ at $x \to x_0$;

(2) if $\alpha(x) \sim \beta(x)$, then $\beta(x) \sim \alpha(x)$ at $x \to x_0$;

(3) if $\alpha(x) \sim \beta(x)$ and $\beta(x) \sim \varphi(x)$, then $\alpha(x) \sim \gamma(x)$ at $x \to x_0$.

Thus the equivalent relation is a reflexive, symmetric and transitive relation.

Examples of equivalent infinitesimals. In the previous sections we have proved that

$$\lim_{x \to 0} \frac{\sin x}{x} = 1 \Rightarrow \sin x \sim x,\ x \to 0$$

and

$$\lim_{x \to 0} \frac{\ln(1+x)}{x} = 1 \Rightarrow \ln(1+x) \sim x,\ x \to 0.$$

It is easy to verify the following:

$$\lim_{x \to 0} \frac{\tan x}{x} = 1 \Rightarrow \tan x \sim x,\ x \to 0,$$

$$\lim_{x \to 0} \frac{\sin^{-1} x}{x} = 1 \Rightarrow \sin^{-1} x \sim x,\ x \to 0,$$

$$\lim_{x \to 0} \frac{\tan^{-1} x}{x} = 1 \Rightarrow \tan^{-1} x \sim x,\ x \to 0.$$

By way of illustration we show that

$$\lim_{x \to 0} \frac{a^x - 1}{x} = \ln a,\ (a > 0,\ a \neq 1).$$

◄ Put $a^x - 1 = y$ so that $y \to 0$ as $x \to 0$, and $a^x = 1 + y$, $x = \dfrac{\ln(1+y)}{\ln a}$. Whence

$$\lim_{x \to 0} \frac{a^x - 1}{x} = \lim_{y \to 0} \frac{y}{\dfrac{\ln(1+y)}{\ln a}} = \lim_{y \to 0} \frac{\ln a}{\dfrac{\ln(1+y)}{y}} = \ln a.$$

Thus $a^x - 1 \sim x \ln a,\ x \to 0$.

In particular, putting $a = \mathrm{e}$, we get

$$\lim_{x\to 0}\frac{\mathrm{e}^x - 1}{x} = 1 \Rightarrow \mathrm{e}^x - 1 \sim x,\ x \to 0.\ \blacktriangleright$$

Let us also verify that $\lim\limits_{x\to 0}\dfrac{(1 + x)^p - 1}{x} = p$ where p is a real number.

◄ Put $(1 + x)^p - 1 = y$ so that $y \to 0$ as $x \to 0$ and $(1 + x)^p = 1 + y$. Whence

$$p \ln (1 + x) = \ln (1 + y).$$

Thus we can write

$$\frac{(1 + x)^p - 1}{x} = \frac{y}{x} = \frac{y}{\ln (1 + y)} \cdot \frac{p \ln (1 + x)}{x}.$$

Working through the limit as $x \to 0$ $(y \to 0)$, we obtain

$$\lim_{x\to 0}\frac{(1 + x)^p - 1}{x} = \lim_{y\to 0}\frac{y}{\ln (1 + y)} \cdot \lim_{x\to 0}\frac{p \ln (1 + x)}{x} = p.$$

Hence, $(1 + x)^p - 1 \sim px,\ x \to 0$. ►

The following list of equivalent infinitesimals is extremely helpful to work through a variety of topics discussed in calculus:

$\sin x \sim x,\ x \to 0;$

$\tan x \sim x,\ x \to 0;$

$\sin^{-1} x \sim x,\ x \to 0;$

$\tan^{-1} x \sim x,\ x \to 0;$

$\ln (1 + x) \sim x,\ x \to 0;$

$a^x - 1 \sim x \ln a,\ x \to 0;$

$\mathrm{e}^x - 1 \sim x,\ x \to 0;$

$(1 + x)^p - 1 \sim px,\ x \to 0.$

Remark. If there exist a nonzero number a and a positive integer m such that $f(x) \sim ax^m$, $x \to 0$ we call ax^m the *principal (power) asymptotic term* of the function $f(x)$ as $x \to 0$. The right-hand terms of asymptotic equalities in the list presented above are the principal (power) terms of their left-hand terms.

Theorem 7.32. *Let $\alpha(x)$, $\beta(x)$, $\alpha_1(x)$ and $\beta_1(x)$ be infinitesimals as $x \to x_0$ and let $\alpha(x) \sim \alpha_1(x)$ and $\beta(x) \sim \beta_1(x)$. If there exists a finite or infinite limit of $\dfrac{\alpha(x)}{\beta(x)}$ at x_0 this limit remains unchanged when $\alpha(x)$ is replaced by $\alpha_1(x)$ and $\beta(x)$ is replaced by $\beta_1(x)$.*

◄ Consider $\frac{\alpha_1(x)}{\beta_1(x)}$ and express it as

$$\frac{\alpha_1(x)}{\beta_1(x)} = \frac{\alpha_1(x)}{\alpha(x)} \cdot \frac{\alpha(x)}{\beta(x)} \cdot \frac{\beta(x)}{\beta_1(x)}. \tag{$*$}$$

By the hypothesis of the theorem

$$\lim_{x \to x_0} \frac{\alpha_1(x)}{\alpha(x)} = 1 \quad \text{and} \quad \lim_{x \to x_0} \frac{\beta(x)}{\beta_1(x)} = 1.$$

Let A be a limit of $\frac{\alpha(x)}{\beta(x)}$ at x_0. Then Theorem 7.20(b) gives

$$\lim_{x \to x_0} \frac{\alpha_1(x)}{\beta_1(x)} = \lim_{x \to x_0} \frac{\alpha_1(x)}{\alpha(x)} \cdot \lim_{x \to x_0} \frac{\alpha(x)}{\beta(x)} \cdot \lim_{x \to x_0} \frac{\beta(x)}{\beta_1(x)} = A.$$

Whence we infer that the ratio $\frac{\alpha_1(x)}{\beta_1(x)}$ also has a limit equal to A at x_0.

If the ratio $\frac{\alpha(x)}{\beta(x)}$ is an infinity as $x \to x_0$ then the right-hand side of ($*$) and, consequently, $\frac{\alpha_1(x)}{\beta_1(x)}$ are also infinities as $x \to x_0$. ►

Example. Compute $\lim_{x \to 0} \frac{1 - \cos x}{\ln(1 + x^2)}$.

◄ Applying Theorem 7.32 and the list of equivalent infinitesimals, we have

$$\lim_{x \to 0} \frac{1 - \cos x}{\ln(1 + x^2)} = \lim_{x \to 0} \frac{2 \sin^2 \frac{x}{2}}{x^2} = \lim_{x \to 0} \frac{2 \frac{x^2}{4}}{x^2} = \frac{1}{2}. \quad ►$$

Theorem 7.33. *For the infinitesimals $\alpha(x)$ and $\beta(x)$ as $x \to x_0$ to be equivalent it is necessary and sufficient that the difference of $\alpha(x)$ and $\beta(x)$ as $x \to x_0$ be an infinitesimal of higher order than $\alpha(x)$ and $\beta(x)$.*

◄ *Necessity.* Let $\alpha(x)$ and $\beta(x)$ be equivalent infinitesimals as $x \to x_0$. We prove that the difference $\gamma(x) = \alpha(x) - \beta(x)$ as $x \to x_0$ is the infinitesimal of higher order than $\beta(x)$ and, consequently, $\alpha(x)$.

Indeed, by the hypothesis of the theorem we have $\alpha(x) \sim \beta(x)$, $x \to x_0$ so that $\lim_{x \to x_0} \frac{\alpha(x)}{\beta(x)} = 1$. Whence

$$\lim_{x \to x_0} \frac{\gamma(x)}{\beta(x)} = \lim_{x \to x_0} \frac{\alpha(x) - \beta(x)}{\beta(x)} = \lim_{x \to x_0} \left(\frac{\alpha(x)}{\beta(x)} - 1 \right) = 0.$$

This means that $\gamma(x)$ is the infinitesimal of higher order than $\beta(x)$ as $x \to x_0$.

Sufficiency. Let $\gamma(x) = \alpha(x) - \beta(x)$ be an infinitesimal of higher order than the infinitesimals $\alpha(x)$ or $\beta(x)$ as $x \to x_0$. We prove that $\alpha(x) \sim \beta(x)$ as $x \to x_0$.

Indeed, as stated $\gamma(x)$ is an infinitesimal of higher order than $\beta(x)$; so

$$\lim_{x \to x_0} \frac{\gamma(x)}{\beta(x)} = 0.$$

Whence

$$\lim_{x \to x_0} \frac{\alpha(x)}{\beta(x)} = \lim_{x \to x_0} \frac{\beta(x) + \gamma(x)}{\beta(x)} = \lim_{x \to x_0} \left(1 + \frac{\gamma(x)}{\beta(x)}\right) = 1,$$

which implies that $\alpha(x)$ and $\beta(x)$ are equivalent infinitesimals as $x \to x_0$. ►

Example. The functions $\alpha(x) = x + 2x^3$ and $\beta(x) = x$ are infinitesimals as $x \to 0$. Then the difference $\gamma(x) = 2x^3$ is an infinitesimal of higher order than $\alpha(x)$ and $\beta(x)$ as $x \to 0$ and, consequently, $\alpha(x) \sim \beta(x)$, $x \to 0$.

We leave it to the reader to solve the following problem.

Problem. Let $\beta_1(x), \beta_2(x), \ldots \beta_m(x)$ be infinitesimals of higher order than $\alpha(x)$ as $x \to x_0$. Prove that the sum $\sigma(x) = \alpha(x) + \beta_1(x) + \beta_2(x) + \ldots + = x\ \beta_m(x)$ is an infinitesimal equivalent to $\alpha(x)$ as $x \to x_0$.

Landau symbols. Let $f(x)$ and $\varphi(x)$ be defined in a neighbourhood Ω of x_0, except probably at x_0, and let $\varphi(x) \neq 0$ in a neighbourhood Ω_0 of x_0, $x \neq x_0$. (Notice that x_0 is a finite or infinite point.)

We say that $f(x)$ is an infinitesimal relative to $\varphi(x)$ as $x \to x_0$ and write $f(x) = o(\varphi(x))$ if

$$\lim_{x \to x_0} \frac{f(x)}{\varphi(x)} = 0.$$

In particular, $f(x) = o(1)$, $x \to x_0$, means that $f(x)$ is an infinitesimal as $x \to x_0$.

Examples. **(1)** $x^2 = o(x)$, $x \to 0$.

(2) $x = o(x^2)$, $x \to \infty$.

In general, $x^\alpha = o(x^\beta)$, $x \to +\infty$, $\alpha < \beta$, and $x^\alpha = o(x^\beta)$, $x \to 0 + 0$, $\alpha > \beta$.

We write $f(x) = O(\varphi(x))$, $x \to x_0$ if there exist a number $M > 0$ and a neighbourhood Ω_0 of x_0 such that

$$|f(x)| \leqslant M|\varphi(x)| \quad \forall x \in \Omega_0, \quad x \neq x_0.$$

Thus $f(x)$ is bounded relative to $\varphi(x)$.

The notation $f(x) = O(1)$, $x \to x_0$ means that the function $f(x)$ is bounded in a neighbourhood of x_0.

Examples. **(1)** $x = O(x^2)$, $[1, +\infty)$.

(2) $x^2 = O(x)$, $[-2, 2]$.

(3) $\sin x = O(1)$, $-\infty < x < +\infty$.

Notice that the symbolic relations we have introduced are not equalities in the ordinary sense. For instance, the "equality" $\sin x = O(1)$, $-\infty < x < +\infty$, does not imply that $O(1) = \sin x$.

The following rules for the symbol "o" can easily be verified:

(a) $o(f(x)) + o(f(x)) = o(f(x))$, $x \to x_0$, $x \in \Omega$;

(b) $o(f(x)) \cdot o(\varphi(x)) = o(f(x) \cdot \varphi(x))$, $x \to x_0$, $x \in \Omega$;

(c) $o(o(f(x))) = o(f(x))$, $x \to x_0$, $x \in \Omega$.

◄ Indeed, let $g_1(x) = o(f(x))$ and $g_2(x) = o(f(x))$. Then (a) becomes $g_1(x) + g_2(x) = o(f(x))$, $x \to x_0$ since

$$\lim_{x \to x_0} \frac{g_1(x) + g_2(x)}{f(x)} = \lim_{x \to x_0} \frac{g_1(x)}{f(x)} + \lim_{x \to x_0} \frac{g(x)}{f(x)} = 0. \blacktriangleright$$

It is also easy to verify the following rules:

(d) $o(f(x)) + O(f(x)) = O(f(x))$, $x \to x_0$;

(e) $o(f(x)) \cdot O(\varphi(x)) = o(f(x) \cdot \varphi(x))$, $x \to x_0$;

(f) $O(f(x)) \cdot O(\varphi(x)) = O(f(x) \cdot \varphi(x))$, $x \to x_0$;

(g) $O(o(f(x)) = o(f(x))$, $o(O(f(x))) = o(f(x))$, $x \to x_0$.

Recall that $f(x)$ and $\varphi(x)$ are equivalent or equal asymptotically as $x \to x_0$ and we write $f(x) \sim \varphi(x)$, $x \to x_0$, if

$$\lim_{x \to x_0} \frac{f(x)}{\varphi(x)} = 1.$$

Using the list of equivalent infinitesimals and Theorem 7.33, we obtain the asymptotic relations

$$\sin x = x + o(x),\ x \to 0;$$

$$e^x = 1 + x + o(x),\ x \to 0;$$

$$\ln(1 + x) = x + o(x),\ x \to 0;$$

$$(1 + x)^\alpha = 1 + \alpha x + o(x),\ x \to 0.$$

The relations of the form $f(x) \sim \varphi(x)$, $f(x) = o(\varphi(x))$ and $f(x) = O(\varphi(x))$ are called the *asymptotic formulas* or *asymptotic relations.*

7.9 Complex Numbers

Representation of a complex number on a plane. A *complex number* is a relation of the form

$$z = x + iy$$

where x and y are arbitrary real numbers and i is the imaginary unity such that $i^2 = -1$.

The numbers x and y are called the *real* and *imaginary parts* of the complex number $z = x + iy$, respectively. We denote the real and imaginary parts of z by

$$x = \operatorname{Re} z \quad \text{and} \quad y = \operatorname{Im} z.$$

The complex number $x + i0$ is thought of as identical to the real number x.

The complex numbers $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ are said to be *equal* ($z_1 = z_2$) if and only if $x_1 = x_2$ and $y_1 = y_2$.

Operations on complex numbers. (a) *Addition of complex numbers.* The sum of $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ is the complex number

$$z = z_1 + z_2 = (x_1 + x_2) + i(y_1 + y_2).$$

The operation of addition of complex numbers obeys

(1) the *commutative law*

$$z_1 + z_2 = z_2 + z_1$$

and

(2) the *associative law*

$$(z_1 + z_2) + z_3 = z_1 + (z_2 + z_3).$$

(b) *Subtraction of complex numbers.* For any complex numbers z_1 and z_2 there exists the number z such that $z_1 = z + z_2$. Then the number z is called the *difference* of z_1 and z_2 and is denoted by

$$z = z_1 - z_2 = (x_1 - x_2) + i(y_1 - y_2).$$

(c) *Multiplication of complex numbers.* The product z_1z_2 of the complex numbers $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ is the complex number

$$z = z_1z_2 = (x_1x_2 - y_1y_2) + i(x_1y_2 + x_2y_1).$$

To memorize this formula it suffices to replace i^2 by -1 in the product $(x_1 + iy_1)(x_2 + iy_2)$.

If z_1 and z_2 are real numbers the operation of multiplication of complex numbers becomes identical to that of real numbers.

It is easy to verify that multiplication of complex numbers obeys

(1) the *commutative law*

$$z_1z_2 = z_2z_1;$$

(2) the *associative law*

$$(z_1z_2)z_3 = z_1(z_2z_3);$$

(3) the *distributive law* (with respect to the operation of addition of complex numbers)

$$(z_1 + z_2)z_3 = z_1 z_3 + z_2 z_3.$$

(d) *Division of complex numbers.* For any complex numbers z_1 and z_2 ($z_2 \neq 0$) there exists a number z such that

$$z_1 = z_2 z. \qquad (*)$$

The complex number z is called the *quotient* of z_1 and z_2 and is denoted by $z = \dfrac{z_1}{z_2}$. We shall derive the quotient formula for z_1 and z_2.

◄ Let $z_1 = x_1 + iy_1$, $z_2 = x_2 + iy_2$ and $z = x + iy$. Then from $(*)$ it follows that

$$x_1 = x_2 x - y_2 y \quad \text{and} \quad y_1 = x_2 y + x y_2.$$

This system of equations is solvable with respect to x and y provided that $z_2 \neq 0$. We have

$$z = x + iy = \frac{x_1 x_2 + y_1 y_2}{x_2^2 + y_2^2} + i\,\frac{x_2 y_1 - x_1 y_2}{x_2^2 + y^2}. \quad ►$$

The complex number

$$\bar{z} = x - iy$$

is called the *complex conjugate* of the complex number $z = x + iy$ and obeys the following rules:

$$\overline{z_1 + z_2} = \bar{z}_1 + \bar{z}_2, \quad \overline{z_1 z_2} = \bar{z}_1 \bar{z}_2,$$

$$\overline{\left(\frac{z_1}{z_2}\right)} = \frac{\bar{z}_1}{\bar{z}_2}, \quad \frac{z_1}{z_2} = \frac{z_1 \bar{z}_2}{z_2 \bar{z}_2},$$

$$z\bar{z} = |z|^2 = x^2 + y^2.$$

Trigonometric and exponential forms of a complex number. The complex number $z = x + iy$ can be represented on the xy-plane by the point M with coordinates (x, y) or by a position vector starting at $O(0, 0)$ and ending at $M(x, y)$ (Fig. 7.35). We shall call the xy-plane the *complex plane.* We also call the x-axis the *real axis* and the y-axis the *imaginary axis.*

It is convenient to use the polar coordinates (r, θ), where r is the length of the vector $\overrightarrow{OM}$ and θ is the angle between the vector $\overrightarrow{OM}$ and the x-axis, to locate the point M in the xy-plane (Fig. 7.36). In polar coordinates

$$x = r\cos\theta, \; y = r\sin\theta.$$

Then the *trigonometric form* of a complex number is

$$z = r(\cos\theta + i\sin\theta),\ z \neq 0$$

with the *absolute value (modulus)* of z

$$r = |z| = \sqrt{x^2 + y^2} = \sqrt{z\bar{z}} \geqslant 0,$$

and the *argument* of z

$$\theta = \operatorname{Arg} z = \arg z + 2k\pi \quad (k = 0,\ \pm 1,\ \pm 2,\ \ldots),$$

where $\arg z$ is the *principal value of the argument.*

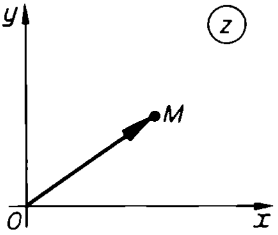


Fig. 7.35

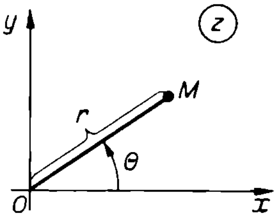


Fig. 7.36

The principal value of the argument of z is defined by

$$-\pi < \arg z \leqslant \pi \quad (\text{or } 0 \leqslant \arg z < 2\pi)$$

or by

$$\arg z = \begin{cases} \tan^{-1}\dfrac{y}{x} & \text{for } x > 0, \\ \pi + \tan^{-1}\dfrac{y}{x} & \text{for } x < 0 \text{ and } y \geqslant 0, \\ -\pi + \tan^{-1}\dfrac{y}{x} & \text{for } x < 0 \text{ and } y < 0, \\ \dfrac{\pi}{2} & \text{for } x = 0 \text{ and } y > 0, \\ -\dfrac{\pi}{2} & \text{for } x = 0 \text{ and } y < 0. \end{cases}$$

The argument of the complex number $z = 0$ is assumed to be undefined and the modulus of $z = 0$ is assumed to be zero.

The complex numbers z_1 and z_2 which are distinct are equal to each other if and only if the moduli of z_1 and z_2 are equal and the arguments of z_1 and z_2 are either equal or differ from each other by a multiple of

2π by a positive integer, i.e.,

$$|z_1| = |z_2| \quad \text{and} \quad \operatorname{Arg} z_1 = \operatorname{Arg} z_2 + 2\pi n,$$

where n is an integer.

Example. Compute the modulus and argument of the complex number

$$z = -\sin\frac{\pi}{8} - i\cos\frac{\pi}{8}.$$

◄ We have

$$x = -\sin\frac{\pi}{8} < 0 \quad \text{and} \quad y = -\cos\frac{\pi}{8} < 0.$$

The principal value of the argument of z is

$$\begin{aligned}\arg z &= -\pi + \tan^{-1}\left(\cot\frac{\pi}{8}\right) = -\pi + \tan^{-1}\left[\tan\left(\frac{\pi}{2} - \frac{\pi}{8}\right)\right] \\ &= -\pi + \tan^{-1}\left(\tan\frac{3\pi}{8}\right) = -\pi + \frac{3\pi}{8} = -\frac{5\pi}{8}.\end{aligned}$$

Hence

$$\operatorname{Arg} z = -\frac{5}{8}\pi + 2k\pi, \ k = 0, \pm 1, \pm 2, \ldots$$

$$|z| = \sqrt{\sin^2\frac{\pi}{8} + \cos^2\frac{\pi}{8}} = 1. \ ►$$

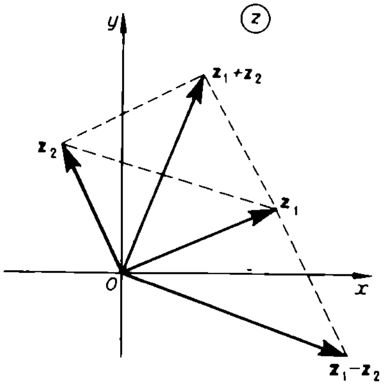


Fig. 7.37

The relationship between the complex numbers and vectors on a plane enables us to interpret the operations of addition and subtraction of complex numbers as vector addition and subtraction as shown in Fig. 7.37.

It is easy to verify the following

$$|z_1 + z_2| \leqslant |z_1| + |z_2|,$$

$$|z_1 - z_2| \geqslant ||z_1| - |z_2||.$$

When dealing with complex numbers it is helpful to use *Euler's formula*

$$\cos\theta + i\sin\theta = e^{i\theta},$$

which enables us to present a complex number z in the *exponential form* as

$$z = re^{i\theta}.$$

It is advisable to present complex numbers in the trigonometric and exponential forms when we wish to multiply or divide complex numbers. If $z_1 = r_1 e^{i\theta_1}$ and $z_2 = r_2 e^{i\theta_2}$ then

$$r_1 e^{i\theta_1} r_2 e^{i(\theta_2)} = r_1 r_2 e^{i(\theta_1 + \theta_2)}.$$

◄ Indeed

$$z_1 z_2 = r_1 e^{i\theta_1} r_2 e^{i\theta_2} = r_1(\cos\theta_1 + i\sin\theta_1) r_2(\cos\theta_2 + i\sin\theta_2)$$
$$= r_1 r_2[\cos\theta_1 \cos\theta_2 - \sin\theta_1 \sin\theta_2) + (\sin\theta_1 \cos\theta_2 + \sin\theta_2 \cos\theta_1)]$$
$$= r_1 r_2(\cos(\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2)) = r_1 r_2\, e^{i(\theta_1 + \theta_2)}.$$

Thus, multiplication of complex numbers involves multiplication of moduli and addition of arguments, viz.,

$$|z_1 z_2| = |z_1|\,|z_2|$$

and

$$\arg(z_1 z_2) = \arg z_1 + \arg z_2. \quad ►$$

Analogously, the operation of division of complex numbers can be described as

$$\frac{z_1}{z_2} = \frac{r_1 e^{i\theta_1}}{r_2 e^{i\theta_2}} = \frac{r_1}{r_2}\, e^{i(\theta_1 - \theta_2)}$$

provided that $r_2 \neq 0$. Whence it follows that

$$\left|\frac{z_1}{z_2}\right| = \frac{|z_1|}{|z_2|}$$

and

$$\arg\frac{z_1}{z_2} = \arg z_1 - \arg z_2.$$

Extracting a root of a complex number. We introduce the operation of raising the complex number z to the nth power as

$$z^n = \underbrace{z\, z\, \ldots\, z}_{n \text{ times}}.$$

Let us represent the complex number z in the exponential and trigonometric forms as

$$z = re^{i\theta} = r(\cos\theta + i\sin\theta).$$

Then we arrive at the formula

$$z^n = (re^{i\theta})^n = r^n e^{in\theta} = r^n(\cos n\theta + i\sin n\theta).$$

Setting $r = 1$, we obtain *de Moivre's formula*

$$(\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta.$$

Based on this experience we can define the operation of extracting a root as follows. The complex number w is called the ***n*th root of the complex number** z if

$$w^n = z.$$

In symbols, we write $w = \sqrt[n]{z}$.

We shall show that for any $z \neq 0$ the root $\sqrt[n]{z}$ assumes n distinct values. Substituting

$$z = re^{i\theta} \text{ and } w = \varrho e^{i\varphi}$$

into $w = \sqrt[n]{z}$, we obtain

$$\varrho^n e^{in\varphi} = re^{i\theta}.$$

Recall that given two equal complex numbers the moduli of these numbers are equal, and the arguments are either equal or differ by a positive multiple of 2π. Then, from the above identity it follows that

$$\varrho^n = r \text{ and } n\varphi = \theta + 2k\pi$$

or

$$\varrho = \sqrt[n]{r} \text{ and } \varphi = \frac{\theta + 2k\pi}{n}. \tag{*}$$

Thus, the moduli of all the nth roots of z are equal and their arguments differ by the multiples of $2\pi/n$ by positive integers. Whence it follows that the nth roots can be displayed on the complex plane as the vertices of the regular polygon of n sides inscribed in the circle with the radius $\sqrt[n]{|z|}$ and centre at the point $z = 0$ (Fig. 7.38).

Substituting $k = 0, 1, 2, \ldots, n - 1$ into (*), we get n distinct complex numbers

$$\sqrt[z]{n} = \sqrt[n]{r}\left(\cos\frac{\theta + 2\pi k}{n} + i\sin\frac{\theta + 2\pi k}{n}\right),$$

$$k = 0, 1, 2, \ldots, n - 1$$

or

$$\sqrt[n]{z} = \sqrt[n]{r}\, e^{i\frac{\theta + 2\pi k}{n}}, \quad k = 0, 1, 2, \ldots, n - 1. \qquad (**)$$

Example. Compute all the values of $\sqrt[3]{i}$.

◀ Writing down $z = i$ in the exponential form, we have

$$z = i = e^{i\frac{\pi}{2}}.$$

Then we obtain from (**)

$$w_k = e^{i\left(\frac{\pi}{6} + \frac{2\pi k}{3}\right)}, \quad k = 0, 1, 2.$$

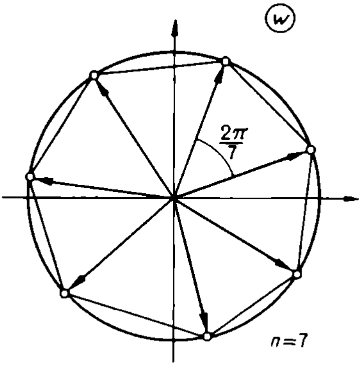


Fig. 7.38

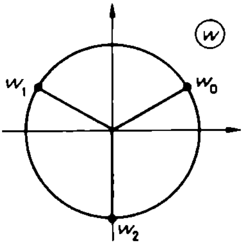


Fig. 7.39

Whence

$$w_0 = e^{i\frac{\pi}{6}} = \cos\frac{\pi}{6} + i\sin\frac{\pi}{6} = \frac{\sqrt{3}}{2} + i\frac{1}{2} = \frac{\sqrt{3} + i}{2},$$

$$w_1 = e^{i\frac{5\pi}{6}} = \cos\frac{5\pi}{6} + i\sin\frac{5\pi}{6}$$

$$= -\frac{\sqrt{3}}{2} + i\frac{1}{2} = \frac{-\sqrt{3} + i}{2},$$

$$w_2 = e^{i\frac{3\pi}{2}} = \cos\frac{3\pi}{2} + i\sin\frac{3\pi}{2} = -i$$

(see Fig. 7.39). ▶

Limits of sequences of complex numbers. Let $\{z_n\}$ be a sequence of complex numbers, where $z_n = x_n + iy_n$. A complex number z is a limit of

$\{z_n\}$ if given any $\varepsilon > 0$ there exists a number $N = N(\varepsilon)$ such that

$$|z_n - z| < \varepsilon$$

whenever $n \geqslant N$.

A sequence $\{z_n\}$ is called *convergent* to a number z if z is the limit of $\{z_n\}$. In symbols, we write

$$z = \lim_{n\to\infty} z_n \text{ or } z_n \to z \ (n \to \infty)$$

to mean that z is a limit of z_n or z_n converges to z, or, using logical symbols,

$$(\lim_{n\to\infty} z_n = z) \Leftrightarrow (\forall\varepsilon > 0 \quad \exists N \quad \forall n \geqslant N,\ |z_n - z| < \varepsilon).$$

Every term $z_n = x_n + iy_n$ of $\{z_n\}$ is associated with the real numbers x_n and y_n. Hence the sequence $\{z_n\}$ of complex numbers corresponds to two sequences $\{x_n\}$ and $\{y_n\}$ of real numbers comprising the real and imaginary parts of terms z_n, respectively.

Theorem 7.34. *A sequence $\{z_n\}$ of complex numbers $z_n = x_n + iy_n$ is convergent if and only if both the sequences $\{x_n\}$ and $\{y_n\}$ of real numbers are convergent.*

◄ Let $\{z_n\}$ converge to a complex number z. This means that for any $\varepsilon > 0$ there exists N such that $|z_n - z| < \varepsilon$ whenever $n \geqslant N$. Since $|x_n - x| \leqslant |z_n - z| < \varepsilon$ and $|y_n - y| \leqslant |z_n - z| < \varepsilon$ then $\lim\limits_{n\to\infty} x_n = x$ and $\lim\limits_{n\to\infty} y_n = y$. Conversely, if $\lim\limits_{n\to\infty} x_n = x$ and $\lim\limits_{n\to\infty} y_n = y$ then for any $\varepsilon > 0$ there exists N_1 such that

$$|x_n - x| < \frac{\varepsilon}{\sqrt{2}} \text{ and } |y_n - y| < \frac{\varepsilon}{\sqrt{2}}.$$

Thus

$$|z_n - z| = |(x_n + iy_n) - (x + iy)| = \sqrt{(x_n - x^2) + (y_n - y)^2} < \varepsilon.$$

Whence $\lim\limits_{x\to\infty} (x_n + iy_n) = x + iy$. ►

By virtue of this theorem all results concerning sequences of real numbers are fully applicable to sequences of complex numbers.

Exercises

1. Prove that the limit of $\left\{\frac{1}{n^2}\right\}$ is equal to zero. Verify what n satisfy the inequality $\frac{1}{n^2} < \varepsilon$ provided that (a) any $\varepsilon > 0$, (b) $\varepsilon = 0.1$ and (c) $\varepsilon = 0.01$.

2. Prove that the limit of $\left\{\frac{n}{n+1}\right\}$ is equal to 1.

Compute the following limits

3. $\lim\limits_{n\to\infty} \dfrac{(n+2)^2}{n^2}$. **4.** $\lim\limits_{n\to\infty} \dfrac{2n^3-n^2+1}{n^4+16n+2}$. **5.** $\lim\limits_{n\to\infty} \dfrac{n^2-n+5}{2n+6}$.

6. $\lim\limits_{n\to\infty} \dfrac{\sqrt[3]{2n^3+n-2}}{n+1}$. **7.** $\lim\limits_{n\to\infty} \dfrac{\sqrt[3]{n^2+6}-n}{\sqrt{n^2+n+1}+\sqrt[5]{n^4+1}}$.

8. $\lim\limits_{n\to\infty} \dfrac{\sqrt[4]{n^5+1}+\sqrt[5]{n^2+2}}{\sqrt[5]{n^4+3}+\sqrt{n^3+5}}$.

(*Hint.* On evaluating a limit of a quotient of polynomials it is helpful to divide the nominator and denominator by n^p where p is the greatest exponent in the denominator. This approach is also used when we evaluate limits of fractions involving irrational functions.)

Compute the following limits

9. $\lim\limits_{n\to\infty} \dfrac{n!}{(n+1)!-n!}$. **10.** $\lim\limits_{n\to\infty} \left(\dfrac{1}{2}+\dfrac{1}{4}+\ldots+\dfrac{1}{2^n}\right)$.

11. $\lim\limits_{n\to\infty} \dfrac{1}{n^2}(1+2+\ldots+n)$. **12.** $\lim\limits_{n\to\infty} \dfrac{n\cos n!}{n^2+1}$. **13.** $\lim\limits_{n\to\infty} (\sqrt{n+2}-\sqrt{n})$. **14.** $\lim\limits_{n\to\infty} (\sqrt{n^2-4n+5}-n)$. **15.** $\lim\limits_{n\to\infty} n(\sqrt{n^2+1}-n)$.

16. $\lim\limits_{x\to 1} \dfrac{x^2-2x+1}{x^3-x}$. **17.** $\lim\limits_{x\to -1} \dfrac{x^2-1}{x^2+3x+2}$. **18.** $\lim\limits_{x\to -2} \dfrac{x^3+3x^2+2x}{x^2-x-6}$.

19. $\lim\limits_{x\to 0} \dfrac{\sqrt{1+x^2}-1}{x}$. **20.** $\lim\limits_{x\to 7} \dfrac{2-\sqrt{x-3}}{x^2-49}$. **21.** $\lim\limits_{x\to 0} \dfrac{\sqrt[3]{1+x^2}-1}{x^2}$.

22. $\lim\limits_{x\to\infty} \dfrac{x^3+x}{x^4-3x^2+1}$. **23.** $\lim\limits_{x\to\infty} \dfrac{1+x-2x^3}{1+x^2+3x^3}$.

Using equivalent infinitesimals, compute the following limits

24. $\lim\limits_{x\to 0} \dfrac{\sin 2x}{x}$. **25.** $\lim\limits_{x\to 0} \dfrac{\sin \alpha x}{\sin \beta x}$ (α, β are constants). **26.** $\lim\limits_{x\to 0} \dfrac{\tan 3x}{\sin 2x}$.

27. $\lim\limits_{x\to 0} \dfrac{\sin^{-1} 3x}{x}$. **28.** $\lim\limits_{x\to 0} \dfrac{\tan^{-1} 2x}{\sin 3x}$. **29.** $\lim\limits_{x\to \pi} \dfrac{\sin 3x}{\sin 2x}$. **30.** $\lim\limits_{x\to 1} \dfrac{1-x^2}{\sin \pi x}$.

31. $\lim\limits_{x\to 0} \dfrac{\ln(1+x^2)}{\sin^2 3x}$. **32.** $\lim\limits_{x\to \mathrm{e}} \dfrac{\ln x-1}{x-\mathrm{e}}$. **33.** $\lim\limits_{x\to 0} \dfrac{\ln\cos x}{x^2}$.

34. $\lim\limits_{x\to 0} \dfrac{\mathrm{e}^{x^2}-1}{\sin^2 x}$. **35.** $\lim\limits_{x\to 0} \dfrac{\ln(1-x)}{\mathrm{e}^{\sin x}-1}$. **36.** $\lim\limits_{x\to 0} \dfrac{\mathrm{e}^{ax}-\mathrm{e}^{bx}}{x}$ (a, b are constants). **37.** $\lim\limits_{x\to\infty} \dfrac{1}{x}(\mathrm{e}^{1/x}-1)$.

Compute the limits

38. $\lim_{x\to\infty}\left(\frac{x}{1+x}\right)^x$. **39.** $\lim_{x\to\infty}\left(\frac{x+1}{x-2}\right)^{3x}$. **40.** $\lim_{x\to 0}(\cos x)^{\frac{1}{x^2}}$.

41. $\lim_{x\to\infty}\left(\frac{x^2+1}{2x^2+1}\right)^{x^2}$.

42. Compute the modulus and the principal value of the argument of the following:

(a) $4+3i$, (b) $-2+2\sqrt{3}i$; (c) $-7-i$;

(d) $-\cos\frac{\pi}{5}+i\sin\frac{\pi}{5}$; (e) $4-3i$.

43. Write down complex numbers in the trigonometric and exponential forms:

(a) -2; (b) $2i$; (c) $-i$; (d) $-\sqrt{2}+i\sqrt{2}$.

44. Compute the following:

(a) $(2-2i)^7$; (b) $(\sqrt{3}-3i)^6$; (c) $\left(\frac{1-i}{1+i}\right)^8$.

45. Compute all the values of the roots:

(a) $\sqrt[4]{1}$; (b) $\sqrt{i}$; (c) $\sqrt[4]{-i}$; (d) $\sqrt[3]{-1+i}$; (e) $\sqrt{2-2\sqrt{3}i}$.

Answers

1. (a) $n > \frac{1}{\sqrt{\varepsilon}}$; (b) $n \geqslant 4$; (c) $n > 10$. **3.** 1. **4.** 0. **5.** ∞. **6.** $\sqrt[3]{2}$. **7.** -1. **8.** 0. **9.** 0. **10.** 1. **11.** 1/2. **12.** 0. **13.** 0. **14.** -2. **15.** 1/2. **16.** 0. **17.** -2. **18.** $-2/5$. **19.** 0. **20.** $-1/56$. **21.** 1/3. **22.** 0. **23.** $-2/3$. **24.** 2. **25.** α/β. **26.** 3/2. **27.** 3. **28.** 2/3. **29.** $-3/2$ (put $\pi - x = t$). **30.** $2/\pi$. **31.** 1/9. **32.** 1/e. **33.** $-1/2$. **34.** 1. **35.** -1. **36.** $a-b$. **37.** 1. **38.** 1/e. **39.** e^9. **40.** $e^{-1/2}$. **41.** 0. **42.** (a) $r=5$, $\theta=\tan^{-1}3/4$; (b) $r=4$, $\theta=\frac{2\pi}{3}$; (c) $r=5\sqrt{2}$, $\theta=-\pi+\tan^{-1}\frac{1}{7}$; (d) $r=1$, $\theta=\frac{4\pi}{5}$; (e) $r=5$, $\theta=-\tan^{-1}\frac{3}{4}$. **43.** (a) $2(\cos\pi+i\sin\pi)$, $2e^{i\pi}$; (b) $2\left(\cos\frac{\pi}{2}+i\sin\frac{\pi}{2}\right)$, $2e^{i\frac{\pi}{2}}$; (c) $\cos\left(-\frac{\pi}{2}\right)+i\sin\left(-\frac{\pi}{2}\right)$, $1\cdot e^{-i\frac{\pi}{2}}$; (d) $2\left(\cos\frac{3\pi}{4}+i\sin\frac{3\pi}{4}\right)$, $2e^{i\frac{3\pi}{4}}$. **44.** (a) $2^{10}(1+i)$; (b) 1728; (c) 1. **45.** (a) ± 1, $\pm i$; (b) $\pm\frac{1}{2}(1+i)$; (c) $\pm\left(\cos\frac{\pi}{8}-i\sin\frac{\pi}{8}\right)$, $\pm\left(\cos\frac{3\pi}{8}+i\sin\frac{3\pi}{8}\right)$; (d) $\frac{\sqrt[3]{4}}{2}(1+i)$, $\sqrt[6]{2}\left(-\cos\frac{\pi}{12}+i\sin\frac{\pi}{12}\right)$, $\sqrt[6]{2}\left(\sin\frac{\pi}{12}-i\cos\frac{\pi}{12}\right)$; (e) $\pm(\sqrt{3}-i)$.

Chapter 8

Differential Calculus. Functions of One Variable

8.1 Derivatives and Differentials

The derivative. Let $y = f(x)$ be a function defined on an open interval (a, b) and let x be a point in (a, b). Consider a point $x + \Delta x$, where Δx is a positive or negative increment of the variable x such that the point $x + \Delta x$ is contained in (a, b). The increment Δx will produce an increment Δy in $y = f(x)$ so that

$$\Delta y = f(x + \Delta x) - f(x).$$

We can write a ratio of two increments Δy and Δx as

$$\frac{\Delta y}{\Delta x} = \frac{f(x + \Delta x) - f(x)}{\Delta x} \quad (\Delta x \neq 0).$$

For a given x this ratio becomes a function of Δx, i.e.,

$$\frac{\Delta y}{\Delta x} = \varphi(\Delta x).$$

Definition. If there exists a limit of the ratio $\frac{\Delta y}{\Delta x}$ as $\Delta x \to 0$ this limit is called the *derivative* of the function $y = f(x)$ at the point x and is denoted by $f'(x)$ or $y'(x)$ or y'_x.

Thus by definition we have

$$f'(x) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}. \qquad (*)$$

Examples. **(1)** Let $y = x^2$. Then for any x and any Δx we have $\Delta y = (x + \Delta x)^2 - x^2 = 2x\Delta x + (\Delta x)^2$. Whence $\frac{\Delta y}{\Delta x} = 2x + \Delta x$ and $\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = 2x$. By definition $\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = y'(x)$. Hence $y = x^2$ has the derivative $y' = 2x$, i.e., $(x^2)' = 2x$ at every point x.

(2) Let $y = e^x$. Then for any x and any Δx we have $\Delta y = e^{x + \Delta x} - e^x = e^x(e^{\Delta x} - 1)$. Whence $\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{e^x(e^{\Delta x} - 1)}{\Delta x} = e^x \lim_{\Delta x \to 0} \frac{e^{\Delta x} - 1}{\Delta x} = e^x$. Thus $(e^x)' = e^x$ for all x.

Remark. Sometimes it is convenient to express (∗) in the following form. Let $f(x)$ be defined at x_0 and in a neighbourhood of x_0. Then

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

provided that this limit exists.

We say that the function $f(x)$ has a *derivative* on (a, b) if this derivative $f'(x)$ exists at each point $x \in (a, b)$.

We leave it to the reader to solve the following problems by applying the definition of the derivative given above.

Problems. (1) Using the definition of the derivative find $f'(0)$ for the function

$$f(x) = \begin{cases} x^2 \sin \dfrac{1}{x} & \text{for} \quad x \neq 0, \\ 0 & \text{for} \quad x = 0. \end{cases}$$

(2) Let $f(x)$ be a periodic function with period T. Prove that if $f(x)$ has a derivative then the derivative of $f(x)$ is also a periodic function with period T.

(3) Prove that the derivative of an even function is also an even function and the derivative of an odd function is an odd function provided that these derivatives exist.

Geometric interpretation of the derivative. We consider the graph of a function $y = f(x)$ which is defined on (a, b) (Fig. 8.1) and choose the points $M(x, f(x))$ and $P(x + \Delta x, f(x + \Delta x))$ on the curve $y = f(x)$. We also draw the line through the points M and P.

Suppose that the point P is moving along the curve $y = f(x)$ towards the point M (or, which is the same, Δx tends to zero) so that the line MP is changing its position until MP coincides with the line MT. The line MT which defines the limiting position of MP as $\Delta x \to 0$ is called the *tangent* to the curve $y = f(x)$ at the point M. Notice that when the point P moves to the point M the angle TMP tends to zero.

From Fig. 8.1 it is easily seen that the slope k of the line MP is

$$k = \tan \alpha = \frac{\Delta y}{\Delta x}.$$

Let $\varphi \neq \pi/2$ be the angle between the tangent MT and the x-axis. Since the slope of the tangent MT to the curve $y = f(x)$ at M is the limit of the slope of MP as the point P moves to the point M along the curve and, consequently, as $\Delta x \to 0$, we get

$$\tan \varphi = \lim_{P \to M} \tan \alpha = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

The latter limit, if it exists, is the derivative $f'(x)$ and is equal to tan φ. Therefore, the derivative $f'(x)$ of the function $y = f(x)$ is the slope of the tangent to the curve $y = f(x)$ at the point specified by the abscissa x.

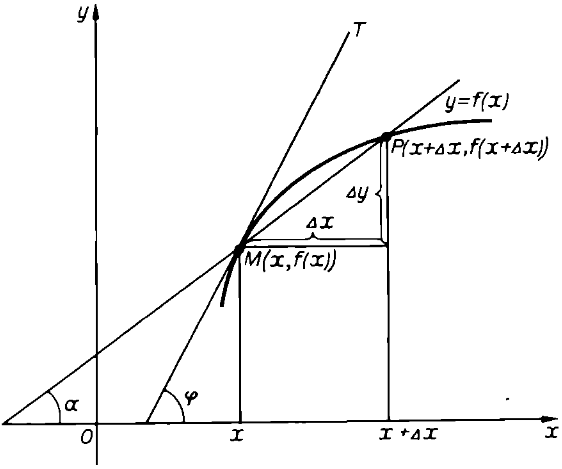


Fig. 8.1

Equations of tangents and normals to a curve. Let a curve be defined by a function $y = f(x)$ and let $M_0(x_0, f(x_0))$ be a point on the curve. We assume that $f(x)$ has a derivative at x_0 and derive the equation of the tangent to the curve at M_0.

The equation of the line through the point $M_0(x_0, y_0)$ is

$$y - y_0 = k(x - x_0),$$

where k is the slope of the line.

The slope k_τ of the tangent to the curve $y = f(x)$ at M_0 is equal to the derivative $f'(x_0)$ so that the equation of the tangent takes the form

$$y - y_0 = f'(x_0)(x - x_0) \quad (y_0 = f(x_0)).$$

The *normal* to the curve at a given point is the line which passes through this point and is perpendicular to the tangent to the curve at this point. This implies that the slope k_n of the normal is related to the slope k_τ of the tangent as

$$k_n = -\frac{1}{k_\tau} \quad \text{or} \quad k_n = -\frac{1}{f'(x_0)} \quad (f'(x_0) \neq 0)$$

so that the equation of the normal to the curve $y = f(x)$ at $M_0(x_0, y_0)$ is

$$y - y_0 = -\frac{1}{f'(x_0)}(x - x_0) \quad (f'(x_0) \neq 0).$$

When $f'(x_0) = 0$ the equation of the normal becomes $x = x_0$.

Example. Write down the equations of the tangent and normal to the curve $y = x^2$ at the point $O(0, 0)$.

◄ We have $f(x) = x^2$, $f'(x) = 2x$ and $f'(0) = 0$. Then the equation of the tangent is $y - 0 = 0(x - 0)$ or $y = 0$, i.e., the tangent coincides with the x-axis, and the equation of the normal is $x = 0$, i.e., the normal coincides with the y-axis (Fig. 8.2). ►

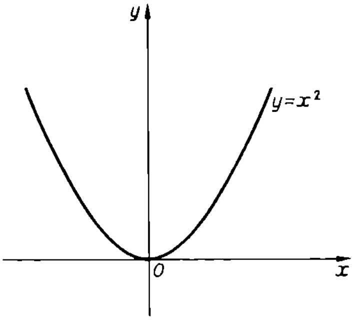


Fig. 8.2

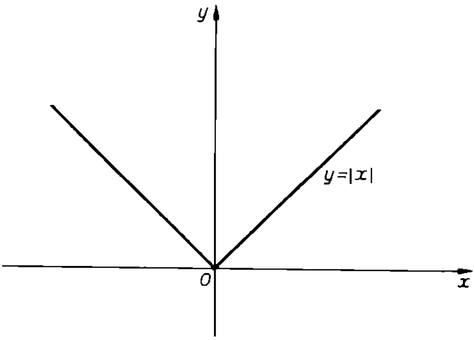


Fig. 8.3

Application of the derivative in mechanics. Let $s = s(t)$ be an equation (a law) of rectilinear motion of a point, which specifies the distance travelled by the point as a function of time t. Let Δs be the distance travelled by the point during the time interval Δt from t to $t + \Delta t$, i.e.,

$$\Delta s = s(t + \Delta t) - s(t).$$

The ratio $\Delta s/\Delta t$ is called the *mean velocity* in the time interval Δt. We define the *velocity* v of the point at the moment t as the limit of the mean velocity in the time interval Δt as $\Delta t \to 0$ and write

$$v(t) = \lim_{\Delta t \to 0} \frac{\Delta s}{\Delta t} = s'(t).$$

Hence the velocity $v(t)$ is equal to the derivative of the distance s with respect to time t, i.e., $v(t) = s'(t)$.

Example. Let us consider the law of rectilinear motion $s = t^2$ where the distance s is measured in metres and the time t is measured in seconds and compute the velocity at $t = 3$ s.

◄ The velocity of a point at any t is given by

$$v = \frac{ds}{dt} = 2t.$$

Whence we have $v = 6\,\mathrm{m/s}$ at $t = 3\,\mathrm{s}$. ►

Right-hand and left-hand derivatives. The *right-hand derivative* $f'(x + 0)$ of the function $y = f(x)$ at a point x is

$$f'(x + 0) = \lim_{\substack{\Delta x \to 0 \\ \Delta x > 0}} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0 + 0} \frac{\Delta y}{\Delta x}$$

and the *left-hand derivative* of $y = f(x)$ at x is

$$f'(x - 0) = \lim_{\substack{\Delta x \to 0 \\ \Delta x < 0}} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0 - 0} \frac{\Delta y}{\Delta x}$$

provided the limits involved exist.

It is easy to see that for the derivative $f'(x)$ to exist at a point x it is necessary and sufficient that the function $y = f(x)$ have the right-hand and left-hand derivatives at x and these derivatives be equal so that

$$f'(x + 0) = f'(x - 0) = f'(x).$$

By way of illustration we shall show that there exist functions which have right-hand and left-hand derivatives at x but have no derivative at x. Consider the function $f(x) = |x|$. The ratio

$$\frac{f(0 + \Delta x) - f(0)}{\Delta x} = \frac{|\Delta x|}{\Delta x}$$

is equal to 1 for $\Delta x > 0$ and is equal to -1 for $\Delta x < 0$. Hence $f(x) = |x|$ has the right-hand derivative $f'(0 + 0) = \lim\limits_{\substack{\Delta x \to 0 \\ \Delta x > 0}} \frac{|\Delta x|}{\Delta x} = 1$ and the left-hand derivative $f'(0 - 0) = \lim\limits_{\substack{\Delta x \to 0 \\ \Delta x < 0}} \frac{|\Delta x|}{\Delta x} = -1$ at $x = 0$ but these derivatives are distinct. Consequently, $f(x) = |x|$ has no derivative at $x = 0$. Geometrically this means that there is no tangent to the curve $y = |x|$ at the point $O(0, 0)$ (Fig. 8.3).

Let $f(x)$ be continuous at a point x_0. We say that $f(x)$ has the *infinite derivative* equal to $+\infty$ or to $-\infty$ at x_0 if at this point

$$f'(x_0) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = +\infty \quad \text{or} \quad f'(x_0) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = -\infty,$$

respectively.

This implies that the tangent to the curve $y = f(x)$ at the point $(x_c, f(x_0))$ is perpendicular to the x-axis. For example, if $f(x) = \sqrt[3]{x}$ then at $x = 0$ we have

$$\frac{\Delta y}{\Delta x} = \frac{f(0 + \Delta x) - f(0)}{\Delta x} = \frac{\sqrt[3]{\Delta x}}{\Delta x} = \frac{1}{\sqrt[3]{(\Delta x)^2}} .$$

Whence it is easy to see that $\Delta y/\Delta x$ tends to $+\infty$ as Δx tends to zero in an arbitrary manner. The tangent to the curve $y = \sqrt[3]{x}$ at the point $O(0, 0)$ coincides with the y-axis (Fig. 8.4).

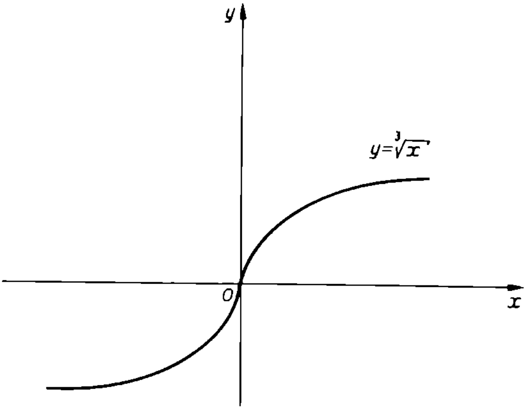


Fig. 8.4

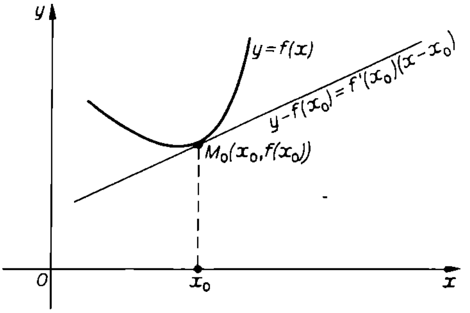


Fig. 8.5

Thus, if a function $f(x)$ has a finite derivative at x_0 then there exists a tangent to the graph of $y = f(x)$ at $M_0(x_0, f(x_0))$ (Fig. 8.5) and this tangent is given by the equation

$$y - f(x_0) = f'(x_0)(x - x_0).$$

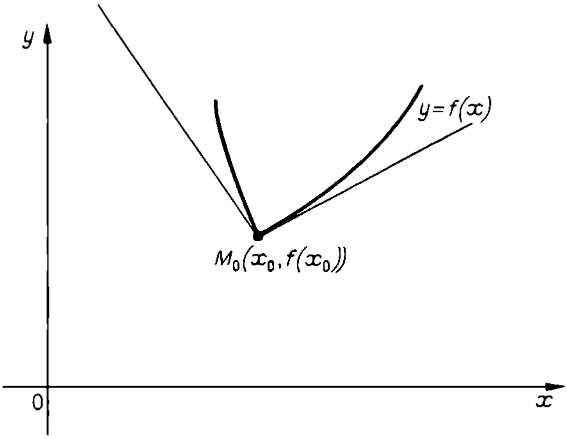


Fig. 8.6

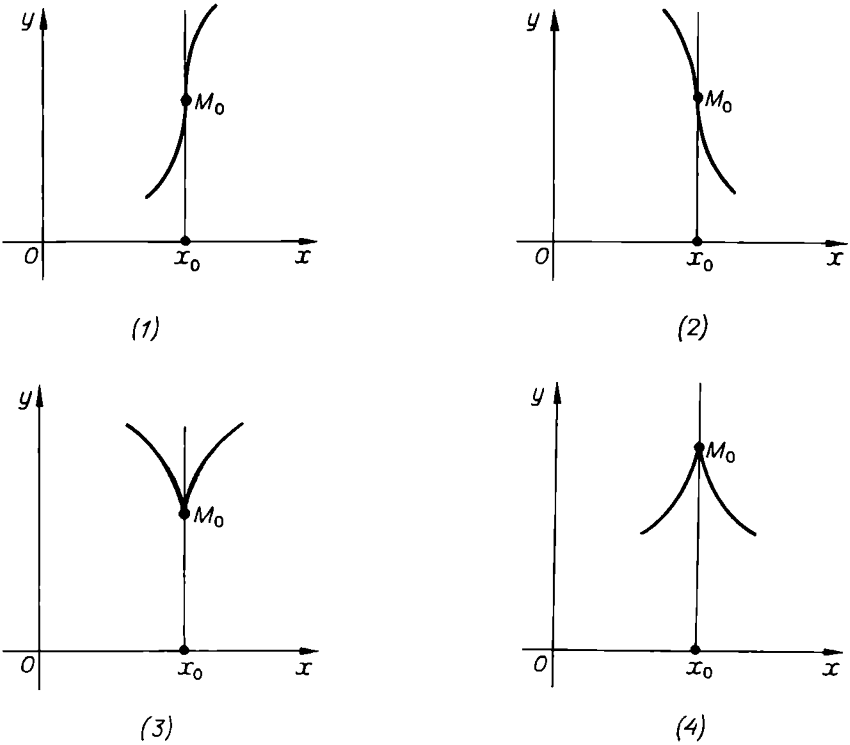


Fig. 8.7

We say that a function $f(x)$ is *smooth* on an open interval (a, b) if $f(x)$ and its derivative are continuous on (a, b). A curve corresponding to a smooth function is also called *smooth* on (a, b).

If a function $y = f(x)$ is continuous at x_0 and has the right- and left-hand derivatives $f'(x_0 + 0)$ and $f'(x_0 - 0)$ and if $f'(x_0 + 0) \neq f'(x_0 - 0)$, then at the point $M_0(x_0, f(x_0))$ there exists no tangent to the curve $y = f(x)$. In this case the curve is not smooth and at $M_0(x_0, f(x_0))$ there exist two lines one of which is a tangent to the left branch of the curve and the other is a tangent to the right branch. The point $M_0(x_0, f(x_0))$ is said to be the *corner point* of the curve (Fig. 8.6). Notice that the point $O(0, 0)$ is the corner point of the function (curve) $y = |x|$.

If a function $f(x)$ is continuous at x_0 and its derivative is infinite at x_0 the following cases should be distinguished

(1) $f'(x_0) = +\infty$;
(2) $f'(x_0) = -\infty$;
(3) $f'(x_0 - 0) = -\infty, \quad f'(x_0 + 0) = +\infty$;
(4) $f'(x_0 - 0) = +\infty, \quad f'(x_0 + 0) = -\infty$.

Figure 8.7 displays the graphs of $y = f(x)$ and tangents $x = x_0$ corresponding to cases (1)-(4).

Differentiable functions. Let $y = f(x)$ be defined on an open interval (a, b) and let x be a point in (a, b). Consider an increment Δx in x such that the point $x + \Delta x$ is contained in (a, b). The increment Δx produces the increment Δy in $f(x)$ so that

$$\Delta y = f(x + \Delta x) - f(x).$$

The function $f(x)$ is called *differentiable* at the point $x \in (a, b)$ if the increment of $f(x)$

$$\Delta y = f(x + \Delta x) - f(x)$$

corresponding to the increment Δx admits a representation of the form

$$\Delta y = A\,\Delta x + \alpha(\Delta x)\,\Delta x,$$

where A is a number independent of Δx (though in general A depends on x) and $\alpha(\Delta x) \to 0$ as $\Delta x \to 0$.

Example. Let $y = x^2$. For any x and any Δx we have

$$\Delta y = (x + \Delta x)^2 - x^2 = \underbrace{2x}_{A}\Delta x + \underbrace{\Delta x}_{\alpha}\Delta x.$$

By definition this means that $y = x^2$ is differentiable at any point x and $A = 2x$, $\alpha(\Delta x) = \Delta x$.

The following theorem specifies the necessary and sufficient conditions for a function to be differentiable.

Theorem 8.1. *For a function $y = f(x)$ to be differentiable at any point x it is necessary and sufficient that $f(x)$ has a finite derivative $f'(x)$ at x.*

◄ *Necessity.* Let $f(x)$ be differentiable at x. We prove that there exists a derivative $f'(x)$ at x. Indeed, since $y = f(x)$ is differentiable at x the increment Δx in x gives rise to the increment Δy which can be expressed as

$$\Delta y = A\,\Delta x + \alpha(\Delta x)\,\Delta x.$$

Whence

$$\frac{\Delta y}{\Delta x} = A + \alpha(\Delta x),$$

where A is independent of Δx, i.e., is constant at a given point x and $\alpha(\Delta x) \to 0$ as $\Delta x \to 0$.

Theorem 7.19 implies that

$$A = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = f'(x).$$

Thus a derivative of $f(x)$ at x does exist.

Sufficiency. Let $f(x)$ have a finite derivative $f'(x)$ at x. We prove that $f(x)$ is differentiable at x. Indeed, since $f'(x)$ exists at x there exists a limit of $\frac{\Delta y}{\Delta x}$ as $\Delta x \to 0$ so that

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = f'(x).$$

Whence, by virtue of Theorem 7.19 it follows that

$$\frac{\Delta y}{\Delta x} = f'(x) + \alpha(\Delta x),$$

where $\alpha(\Delta x) \to 0$ as $\Delta x \to 0$ and, consequently,

$$\Delta y = f'(x)\,\Delta x + \alpha(\Delta x)\,\Delta x. \tag{$*$}$$

Since $f'(x)$ is independent of Δx and $\alpha(\Delta x) \to 0$ as $\Delta x \to 0$ from ($*$) we infer that $y = f(x)$ is differentiable at x. ►

Theorem 8.1 establishes a one-to-one correspondence between the notion of a function differentiable at a point and the notion of a function having a finite derivative at this point, that is, if $f(x)$ is differentiable at a point then it has a finite derivative at this point and vice versa. So the operation of computing a derivative of a function is also called the differentiation of a function.

Continuity of differentiable functions. We shall prove the theorem which establishes a relationship between continuous functions and differentiable functions.

Theorem 8.2. *If a function $y = f(x)$ is differentiable at a point x then $f(x)$ is continuous at x.*

◄ Indeed, if $f(x)$ is differentiable at x then the increment Δy corresponding to the increment Δx in x admits a representation of the form

$$\Delta y = A\,\Delta x + \alpha(\Delta x)\,\Delta x,$$

where A is constant at x and $\alpha(\Delta x) \to 0$ as $\Delta x \to 0$. Whence $\lim\limits_{\Delta x \to 0} \Delta y = 0$ so that $f(x)$ is continuous at x. ►

The converse is not true, namely, if $f(x)$ is continuous at x it is not necessarily differentiable at x. For example, $f(x) = |x|$ is continuous at $x = 0$ but has no derivative at $x = 0$. Hence, $f(x)$ is not differentiable at $x = 0$.

Example. The function

$$f(x) = \begin{cases} x \sin\dfrac{1}{x} & \text{for} \quad x \neq 0, \\ 0 & \text{for} \quad x = 0 \end{cases}$$

is continuous on $(-\infty, +\infty)$. For all $x \neq 0$ $f(x)$ has a derivative; however $f(x)$ has neither right- nor left-hand derivatives at $x = 0$ since

$$\frac{\Delta x \sin\dfrac{1}{\Delta x}}{\Delta x} = \sin\frac{1}{\Delta x}$$

has no limit as $\Delta x \to 0 + 0$ and as $\Delta x \to 0 - 0$.

The examples we have considered involve functions each of which has no derivative only at one point in its domain. The idea that a function may have no derivatives only at a finite number of points prevailed in the eighteenth and early nineteenth centuries. However, later on various eminent mathematicians offered examples of functions continuous on a closed interval $[a, b]$ and having no derivatives at any point of $[a, b]$.

The differential. Let $y = f(x)$ be differentiable at a point x, i.e., let an increment Δx in x produce an increment Δy in $f(x)$ so that

$$\Delta y = A\,\Delta x + \alpha(\Delta x)\,\Delta x,$$

where $\alpha(\Delta x) \to 0$ as $\Delta x \to 0$.

If $y = f(x)$ is differentiable at x the linear part $A\,\Delta x$ of Δy, provided that $A \neq 0$, is called the *differential* of $y = f(x)$ and is denoted by dy or by $df(x)$ so that

$$dy = A\,\Delta x.$$

When $A \neq 0$ we say that $A\,\Delta x$ is the ***principal linear part*** of Δy since $\alpha(\Delta x)\,\Delta x$ is an infinitesimal of higher order than $A\,\Delta x$ as $\Delta x \to 0$. If $A = 0$ the differential is equal to zero.

By virtue of Theorem 8.1 we have $A = f'(x)$. Then the differential dy becomes

$$dy = f'(x)\,\Delta x.$$

Along with the notion of the differential of a function we can introduce the notion of the differential of an independent variable x by putting $dx = \Delta x$. Then the differential of $y = f(x)$ can be written as

$$dy = f'(x)\,dx.$$

Whence we have the *Leibniz* notation for the derivative: $f'(x) = \dfrac{dy}{dx}$. Thus a derivative of $f(x)$ can also be thought of as a quotient of two differentials dy and dx.

We say that a function $y = f(x)$ is ***differentiable on an open interval*** (a, b) if $f(x)$ is differentiable at every point of (a, b).

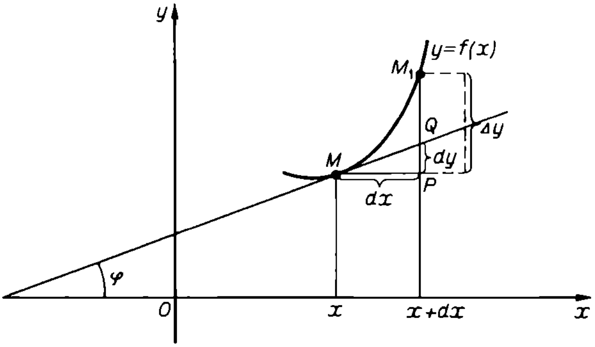


Fig. 8.8

To interpret the notion of a differential geometrically we use Fig. 8.8. Let a function $y = f(x)$ be differentiable on some open interval (a, b). We draw a tangent to the curve $f(x)$ at a point M and choose a point M_1 whose abscissa is $x + dx$. Clearly, $f'(x) = \tan\varphi$. Consider the triangle MPQ. It is easy to see that

$$PQ = MP\tan\varphi = f'(x)dx = dy.$$

Thus, the differential $dy = f'(x)\,dx$ of the function $y = f(x)$ is the increment of the ordinate of the tangent to the curve $y = f(x)$ at M when x is given the increment dx.

8.2 Differentiation Rules

The derivative of a constant function. If $y = C = \text{const}$ at every point of (a, b) then $y' = 0$ in (a, b).

◀ Indeed, if $y = C$ $\forall x \in (a, b)$ then for any $x \in (a, b)$ and any Δx such that $x + \Delta x \in (a, b)$ we have

$$\Delta y = C - C = 0 \quad \text{and} \quad \frac{\Delta y}{\Delta x} = 0 \quad (\Delta x \neq 0).$$

Whence

$$y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = 0 \quad \forall x \in (a, b).$$

Thus

$$(C)' = 0 \quad \text{and} \quad dC = 0. \ \blacktriangleright$$

The derivative of a sum of functions. Let $u(x)$ and $v(x)$ be differentiable at x. Then the sum $y(x) = u(x) + v(x)$ is also differentiable at x and

$$y'(x) = (u(x) + v(x))' = u'(x) + v'(x).$$

◀ Indeed, the increment Δx in x produces the increments in $u(x)$ and in $v(x)$ so that

$$u(x + \Delta x) = u(x) + \Delta u \quad \text{and} \quad v(x + \Delta x) = v(x) + \Delta v.$$

Then the increment Δy in $y = u(x) + v(x)$ becomes

$$\Delta y = [(u + \Delta u) + (v + \Delta v)] - (u + v) = \Delta u + \Delta v.$$

Whence

$$\frac{\Delta y}{\Delta x} = \frac{\Delta u}{\Delta x} + \frac{\Delta v}{\Delta x}. \tag{*}$$

Since $u(x)$ and $v(x)$ are differentiable at x there exist the derivatives $u'(x)$ and $v'(x)$ at x. Then each summand in (*) has a limit as $\Delta x \to 0$ so that there exists a limit of the right-hand side of (*) equal to $u'(x) + v'(x)$. Hence, the left-hand side of (*) also has a limit as $\Delta x \to 0$. In other words, there exists

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = y'(x).$$

Thus, evaluating the limit of (*) as $\Delta x \to 0$, we obtain

$$y'(x) = (u(x) + v(x))' = u'(x) + v'(x)$$

and

$$d(u + v) = du + dv. \ \blacktriangleright$$

Analogously, we prove that

$$(u(x) - v(x))' = u'(x) - v'(x)$$

and

$$d(u - v) = du - dv.$$

We can easily extend these results to any finite number of differentiable functions.

Example. Find the derivative of $y = x^2 + e^x + 2$.

◄ $$y' = (x^2 + e^x + 2)' = (x^2)' + (e^x)' + (2)' = 2x + e^x.$$ ►

The derivative of product of functions (product rule). If $u(x)$ and $v(x)$ are differentiable at x so is the product $u(x)v(x)$ and

$$(u(x)v(x))' = u'(x)v(x) + u(x)v'(x).$$

◄ Let Δx be an increment in x. Then $u(x)$ and $v(x)$ receive the increments Δu and Δv which give rise to the increment

$$\Delta y = (u + \Delta u)(v + \Delta v) - uv = v\,\Delta u + u\,\Delta v + \Delta u\,\Delta v.$$

Consider the ratio

$$\frac{\Delta y}{\Delta x} = v\frac{\Delta u}{\Delta x} + u\frac{\Delta v}{\Delta x} + \Delta u\frac{\Delta v}{\Delta x}. \tag{$*$}$$

We prove that each factor in the right-hand side of $(*)$ has a limit as $\Delta x \to 0$. Indeed, at a given point x $u(x)$ and $v(x)$ are constant. Since $u(x)$ and $v(x)$ are differentiable at x there exist the derivatives

$$u'(x) = \lim_{\Delta x \to 0}\frac{\Delta u}{\Delta x} \quad \text{and} \quad v'(x) = \lim_{\Delta x \to 0}\frac{\Delta v}{\Delta x}.$$

Since $u(x)$ and $v(x)$ are differentiable at x they are continuous at x so that Δu and Δv tend to zero as $\Delta x \to 0$. Thus the right-hand side of $(*)$ has a limit equal to $v(x)u'(x) + u(x)v'(x)$ as $\Delta x \to 0$. Hence there exists a limit of the left-hand side of $(*)$, i.e., there exists $\lim\limits_{\Delta x \to 0}\frac{\Delta y}{\Delta x} = y'(x)$. Evaluating the limit of $(*)$ as $\Delta x \to 0$, we obtain

$$y'(x) = (u(x)v(x))' = u'(x)v(x) + u(x)v'(x).$$

Whence

$$d(uv) = v\,du + u\,dv.$$ ►

Example. Find the derivative of $y = (x^2 - 1)(e^x + 2)$.

◄ $$\begin{aligned} y' &= ((x^2 - 1)(e^x + 2))' = (x^2 - 1)'(e^x + 2) + (x^2 - 1)(e^x + 2)' \\ &= 2x(e^x + 2) + (x^2 - 1)e^x = (x^2 + 2x - 1)e^x + 4x. \end{aligned}$$ ►

Corollary. If a function is multiplied by a constant factor its derivative (or differential) is multiplied by this factor, i.e.,

$$(Cu(x))' = Cu'(x) \quad \text{and} \quad d(Cu(x)) = Cdu \quad (C = \text{const}).$$

The product rule is easily generalized to any finite number of differentiable functions so that

$$(u_1(x)u_2(x) \ldots u_n(x))' = u_1'(x)u_2(x) \ldots u_n(x) + u_1(x)u_2'(x) \ldots u_n(x) + \ldots + u_1(x)u_2(x) \ldots u_n'(x).$$

The derivative of a quotient of functions (quotient rule). If $u(x)$ and $v(x)$ are differentiable at a point x and if $v(x) \neq 0$ at x then the quotient $y(x) = \frac{u(x)}{v(x)}$ is differentiable at x and

$$\left.\begin{aligned} y' &= \left(\frac{u}{v}\right)' = \frac{u'v - uv'}{v^2}, \\ d\left(\frac{u}{v}\right) &= \frac{v\,du - u\,dv}{v^2}, \end{aligned}\right\} \quad (v(x) \neq 0).$$

◄ Since $v(x)$ is differentiable at x, $v(x)$ is continuous at x and, by virtue of Theorem 7.22, $v(x + \Delta x) \neq 0$ for all sufficiently small $|\Delta x|$. Then the ratio

$$\frac{u(x + \Delta x)}{v(x + \Delta x)} = \frac{u + \Delta u}{v + \Delta v}$$

is defined for all sufficiently small $|\Delta x|$.

The increment Δx in x produces the increment

$$\Delta y = \frac{u + \Delta u}{v + \Delta v} - \frac{u}{v} = \frac{v\,\Delta u - u\,\Delta v}{v^2 + v\,\Delta v}.$$

Whence

$$\frac{\Delta y}{\Delta x} = \frac{v\dfrac{\Delta u}{\Delta x} - u\dfrac{\Delta v}{\Delta x}}{v^2 + v\,\Delta v}. \qquad (*)$$

Recall that $u(x)$ and $v(x)$ are differentiable at x. This implies that there exist $\lim\limits_{\Delta x \to 0} \frac{\Delta u}{\Delta x} = u'(x)$ and $\lim\limits_{\Delta x \to 0} \frac{\Delta v}{\Delta x} = v'(x)$ and $\Delta v \to 0$ as $\Delta x \to 0$. Also, at a given point x the values of u and v are constant and, as stated, $v(x) \neq 0$. The right-hand side of $(*)$ thus has a limit as $\Delta x \to 0$ equal to $\frac{vu' - uv'}{v^2}$.

Hence there exists a limit on the left of (∗), i.e., there exists $\lim\limits_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = y'(x)$. Evaluating the limit of (∗) as $\Delta x \to 0$, we obtain

$$y'(x) = \left(\frac{u(x)}{v(x)}\right)' = \frac{u'(x)v(x) - u(x)v'(x)}{v^2(x)},$$

$$(v(x) \neq 0).$$ ►

$$d\left(\frac{u}{v}\right) = \frac{vdu - udv}{v^2},$$

Example. Find the derivative of $y = \frac{e^x - 1}{x^2 + 5}$.

◄ $$y' = \left(\frac{e^x - 1}{x^2 + 5}\right)' = \frac{(e^x - 1)'(x^2 + 5) - (e^x - 1)(x^2 + 5)'}{(x^2 + 5)^2}$$

$$= \frac{e^x(x^2 + 5) - (e^x - 1)2x}{(x^2 + 5)^2} = \frac{(x^2 - 2x + 5)e^x + 2x}{(x^2 + 5)^2}.$$ ►

Problems. (1) Determine whether $f(x) + \varphi(x)$ is differentiable at a point x or not, provided that $f(x)$ is differentiable at x and $\varphi(x)$ is not.

(2) Suppose that $f(x)$ is differentiable at x_0, $f(x_0) \neq 0$, and $\varphi(x)$ is not differentiable at x_0. Prove that $f(x)\varphi(x)$ is not differentiable at x_0.

(3) Suppose that $f(x)$ and $\varphi(x)$ have no derivatives at x_0. Does it imply that (a) $f(x) + \varphi(x)$, (b) $f(x)\varphi(x)$ and (c) $\frac{f(x)}{\varphi(x)}$ have no derivatives at x_0? (Consider the following functions: (i) $f(x) = |x|$, $\varphi(x) = -|x|$, $x_0 = 0$; (ii) $f(x) = \varphi(x) = |x|$, $x_0 = 0$; (iii) $f(x) = \varphi(x) = |x| + 1$, $x_0 = 0$.)

Derivatives of some elementary functions. (1) *Exponential function* $y = a^x$ $(a > 0, a \neq 1)$. This function is defined at every point of the number line. Hence, for any x and any Δx we have

$$\Delta y = a^{x + \Delta x} - a^x = a^x(a^{\Delta x} - 1)$$

and, for $\Delta x \neq 0$,

$$\frac{\Delta y}{\Delta x} = a^x \frac{a^{\Delta x} - 1}{\Delta x},$$

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} a^x \frac{a^{\Delta x} - 1}{\Delta x} = a^x \lim_{\Delta x \to 0} \frac{a^{\Delta x} - 1}{\Delta x} = a^x \ln a.$$

Thus

$$(a^x)' = a^x \ln a.$$

In particular, if $a = \mathrm{e}$ then

$$(\mathrm{e}^x)' = \mathrm{e}^x.$$

(2) *Logarithmic function* $y = \ln x$ $(x > 0)$. For any $x > 0$ and any Δx such that $x + \Delta x > 0$ we have

$$\Delta y = \ln(x + \Delta x) - \ln x = \ln\left(1 + \frac{\Delta x}{x}\right).$$

Whence

$$\frac{\Delta y}{\Delta x} = \frac{\ln\left(1 + \frac{\Delta x}{x}\right)}{\Delta x}$$

and

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{\ln\left(1 + \frac{\Delta x}{x}\right)}{\Delta x} = \lim_{\Delta x \to 0} \frac{\frac{\Delta x}{x}}{\Delta x} = \frac{1}{x}.$$

Thus

$$(\ln x)' = \frac{1}{x}.$$

Using the identity

$$\log_a x = \log_a \mathrm{e} \ln x \quad (a > 0,\ a \neq 1)$$

we arrive at

$$(\log_a x)' = \log_a \mathrm{e}\,(\ln x)' = \frac{\log_a \mathrm{e}}{x} = \frac{1}{x \ln a}$$

and, finally,

$$(\log_a x)' = \frac{1}{x \ln a}.$$

(3) *Power function* $y = x^\alpha$ (α is an arbitrary real number). This function is defined at least for all $x > 0$. Then

$$\Delta y = (x + \Delta x)^\alpha - x^\alpha = x^\alpha \left[\left(1 + \frac{\Delta x}{x}\right)^\alpha - 1\right]$$

and

$$\frac{\Delta y}{\Delta x} = x^\alpha \frac{\left(1 + \frac{\Delta x}{x}\right)^\alpha - 1}{\Delta x}.$$

Recall that $\left(1 + \frac{\Delta x}{x}\right)^\alpha - 1 \sim \alpha \frac{\Delta x}{x}$ as $\Delta x \to 0$ (see Chap. 7). Then

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = x^\alpha \lim_{\Delta x \to 0} \frac{\left(1 + \frac{\Delta x}{x}\right)^\alpha - 1}{\Delta x} = x^\alpha \lim_{\Delta x \to 0} \frac{\frac{\alpha \Delta x}{x}}{\Delta x} = \alpha x^{\alpha - 1}.$$

and

$$(x^{\alpha})' = \alpha x^{\alpha-1}.$$

(4) *Trigonometric functions.* Consider $y = \sin x$, $-\infty < x < +\infty$. For any x and any Δx

$$\Delta y = \sin(x + \Delta x) - \sin x = 2 \sin \frac{\Delta x}{2} \cos\left(x + \frac{\Delta x}{2}\right)$$

and

$$\frac{\Delta y}{\Delta x} = \frac{2 \sin \dfrac{\Delta x}{2}}{\Delta x} \cos\left(x + \frac{\Delta x}{2}\right) = \frac{\sin \dfrac{\Delta x}{2}}{\dfrac{\Delta x}{2}} \cos\left(x + \frac{\Delta x}{2}\right).$$

Since $y = \cos x$ is a continuous function at any point x, $\lim\limits_{\Delta x \to 0} \cos\left(x + \dfrac{\Delta x}{2}\right) = \cos x$ and $\lim\limits_{\Delta x \to 0} \dfrac{\sin \dfrac{\Delta x}{2}}{\dfrac{\Delta x}{2}} = 1$, we obtain

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \left(\frac{\sin \dfrac{\Delta x}{2}}{\dfrac{\Delta x}{2}} \cos\left(x + \frac{\Delta x}{2}\right) \right) = \cos x.$$

Thus

$$(\sin x)' = \cos x.$$

Analogously

$$(\cos x)' = -\sin x.$$

Using formulas for derivatives of $\sin x$ and $\cos x$, we easily arrive at

$$(\tan x)' = \left(\frac{\sin x}{\cos x}\right)' = \frac{(\sin x)' \cos x - \sin x (\cos x)'}{\cos^2 x} = \frac{1}{\cos^2 x} = \sec^2 x$$

so that

$$(\tan x)' = \frac{1}{\cos^2 x} = \sec^2 x \quad \left(x \neq \frac{\pi}{2} + n\pi,\ n = 0,\ \pm 1,\ \pm 2,\ \ldots\right).$$

Analogously

$$(\cot x)' = -\frac{1}{\sin^2 x} = -\operatorname{cosec}^2 x \quad (x \neq n\pi,\ n = 0,\ \pm 1,\ \pm 2,\ \ldots).$$

8.3 Differentiation of Composite and Inverse Functions

Derivatives of composite functions. First we shall prove the following important theorem.

Theorem 8.3 (chain rule). *Let* $u = \varphi(x)$ *be differentiable at* x_0 *and let* $y = f(u)$ *be differentiable at* $u_0 = \varphi(x_0)$. *Then the composite function* $y = f[\varphi(x)]$ *is differentiable at* x_0 *and*

$$\{f[\varphi(x)]\}'_x|_{x=x_0} = f'(u_0)\varphi'(x_0).$$

◀ Following our usual approach we consider the increment Δu in $u = \varphi(x)$ corresponding to the increment Δx in x_0. The increment Δu produces the increment Δy in $y = f(u)$.

Since $y = f(u)$ is differentiable at u_0 we can write

$$\Delta y = f'(u_0)\Delta u + \alpha(\Delta u)\Delta u, \tag{$*$}$$

where $\alpha(\Delta u) \to 0$ as $\Delta u \to 0$.

The function $\alpha(\Delta u)$ is indeterminate as $\Delta u = 0$. If we set $\alpha(0) = 0$, $\alpha(\Delta u)$ becomes continuous as $\Delta u = 0$.

Dividing both sides of $(*)$ by Δx ($\Delta x \neq 0$), we obtain

$$\frac{\Delta y}{\Delta x} = f'(u_0)\frac{\Delta u}{\Delta x} + \alpha(\Delta u)\frac{\Delta u}{\Delta x}. \tag{$**$}$$

Also, $u = \varphi(x)$ is differentiable at x_0 and, consequently, continuous at x_0. Hence, $\Delta u \to 0$ as $\Delta x \to 0$. This implies that $\alpha(\Delta u)$ tends to zero and $\frac{\Delta u}{\Delta x} \to \varphi'(x_0)$ as $\Delta x \to 0$. Therefore the right-hand side of $(**)$ has a limit equal to $f'(u_0)\varphi'(x_0)$ as $\Delta x \to 0$. So there exists a limit of the left-hand side of $(**)$, i.e., $\lim\limits_{\Delta x \to 0}\frac{\Delta y}{\Delta x}$, which is the derivative of the composite function $y = f[\varphi(x)]$ at x_0 with respect to the variable x.

Evaluating the limit of $(**)$ as $\Delta x \to 0$, we obtain

$$\{f[\varphi(x)]\}'_x|_{x=x_0} = f'(u_0)\varphi'(x_0),$$

where $f'(u_0)$ designates the derivative of $f(u)$ with respect to the variable u at the point $u_0 = \varphi(x_0)$ corresponding to the point x_0 of the variable x.

This identity can be written as

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$$

or

$$y'_x = y'_u u'_x. \blacktriangleright$$

Examples. (1) Find the derivative of $y = e^{\sin x}$.

◀ Here y is a composite function of x which can be expressed as $y = e^{u(x)}$, where $u(x) = \sin x$. Then

$$y'_x = (e^u)'_u u'_x = e^u \cos x = e^{\sin x}\cos x. \blacktriangleright$$

(2) Find the derivative of $y = \ln |x|$, $x \neq 0$.

◄ This function is even and is defined at every point of the number line except at $x = 0$. If $x > 0$ then $|x| = x$ and $\ln |x| = \ln x$ so that

$$y'_x = (\ln x)' = \frac{1}{x} \quad (x > 0).$$

If $x < 0$ then $|x| = -x$ and $\ln |x| = \ln(-x)$.

Set $y = \ln u$ and $u = -x$. Then $y = \ln(-x)$ becomes a composite function. By the chain rule we have

$$y'_x = y'_u u'_x = \frac{1}{u}(-1) = \frac{1}{-x}(-1) = \frac{1}{x}$$

so that for $x < 0$ $y'_x = \frac{1}{x}$.

Thus

$$(\ln |x|)' = \frac{1}{x} \quad (x \neq 0). \quad ►$$

Remark. The chain rule is also applicable when we consider any finite chain of functions. For example, if $y = f(u)$, $u = \varphi(z)$ and $z = \psi(x)$ so that $y = f\{\varphi[\psi(x)]\}$ and if there exist the derivatives f'_u, φ'_z and ψ'_x then

$$y'_x = y'_u u'_z z'_x.$$

Invariance of the form of the differential. If $y = f(u)$ is a differentiable function of an independent variable u then

$$dy = f'(u)du, \qquad (*)$$

where $du = \Delta u$.

Let u be a differentiable function $u = \varphi(u)$ of a variable x. Then we may consider y as a composite function $y = f[\varphi(x)]$ of the variable x. Since x is an independent variable we can express the differential dy of the composite function $y = f[\varphi(x)]$ as

$$dy = \{f[\varphi(x)]\}'_x \, dx.$$

The chain rule yields

$$\{f[\varphi(x)]\}'_x = f'(u)\varphi'(x)$$

and we can write

$$dy = f'(u)\varphi'(x)dx.$$

Since $\varphi'(x)dx = du$ we again get the identity $(*)$

$$dy = f'(u)du.$$

Therefore the differential of a function is expressed by the same formula irrespective of whether the argument of this function is an independent variable or a function of another variable. This property is sometimes called the *invariance of the form of the differential.* It is worth mentioning that in the formula $dy = f'(u)du$ the differential du is equal to an arbitrary increment Δu of an independent variable u or, when $u = \varphi(x)$, $du = \varphi'(x)dx$ is a linear part of the increment of $u = \varphi(x)$ and, in general, is not equal to Δu.

Differentiation of inverse functions. Let $y = f(x)$ be defined on a closed interval $[a, b]$. Suppose that the range of $y = f(x)$ is the closed interval $[\alpha, \beta]$ on the y-axis. Furthermore, let each y in $[\alpha, \beta]$ correspond to the only x in $[a, b]$ so that $f(x) = y$ (Fig. 8.9). Then we can specify the function $x = \varphi(y)$ on $[\alpha, \beta]$ by associating every y in $[\alpha, \beta]$ with x in $[a, b]$ such that $f(x) = y$. The function $x = \varphi(y)$ is called the *inverse function* of $y = f(x)$.

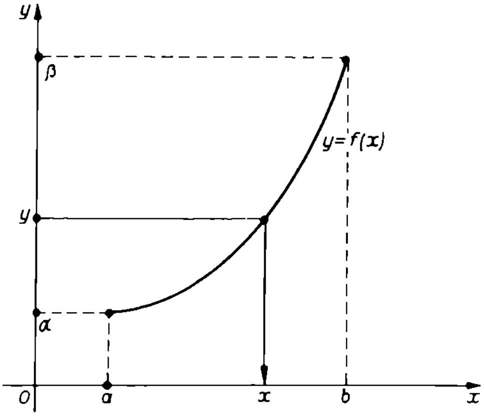


Fig. 8.9

Clearly, if $x = \varphi(y)$ is the inverse of $y = f(x)$ then $y = f(x)$ is the inverse of $x = \varphi(y)$. In this case we speak of $y = f(x)$ and $x = \varphi(x)$ as of *mutually inverse functions.* We can write

$$f[\varphi(y)] = y \quad \text{and} \quad \varphi[f(x)] = x$$

provided that $y = f(x)$ and $x = \varphi(y)$ are mutually inverse functions.

We shall follow a more constructive approach to specification of an inverse function. Let $y = f(x)$ be an equation solvable for x so that every

y is associated with exactly one x. Then we have the equation $x = \varphi(y)$ which defines x as a function of y. The function $x = \varphi(y)$ is the inverse function of $y = f(x)$.

Examples. (1) If $y = 3x$ is defined on [0, 1] then the function $x = y/3$ defined on [0, 3] is the inverse function of y.

(2) The inverse function of $y = x^3$, $-\infty < x < +\infty$, is $x = \sqrt[3]{y}$, $-\infty < y < +\infty$.

(3) Let

$$y = \begin{cases} x & \text{for rational } x, \\ 1 - x & \text{for irrational } x. \end{cases}$$

Then the inverse function of $y(x)$ is

$$x = \begin{cases} y & \text{for rational } y, \\ 1 - y & \text{for irrational } y. \end{cases}$$

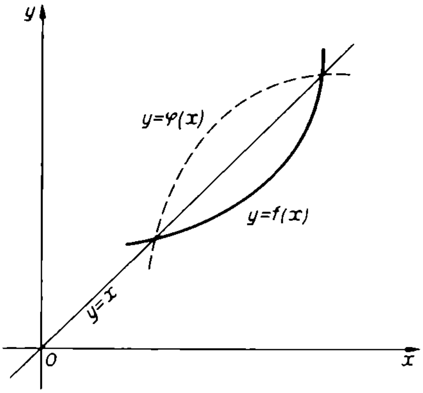


Fig. 8.10

Clearly, the equations $y = f(x)$ and $x = \varphi(y)$ specify the same curve on the xy-plane. If we display independent variables on the x-axis in both cases, i.e., if we plot the functions $y = f(x)$ and $y = \varphi(x)$ instead of $y = f(x)$ and $x = \varphi(y)$, their graphs will be symmetric relative to the line bisecting the first and third quadrants of the coordinate plane (Fig. 8.10).

We say that $y = f(x)$ is an *increasing function* on $[a, b]$ if given any x_1 and x_2 in $[a, b]$ such that $x_1 < x_2$ there holds $f(x_1) < f(x_2)$.

For instance, $f(x) = x^3$ is an increasing function for $-\infty < x < +\infty$.

Theorem 8.4. *Let $y = f(x)$ be a continuous and increasing function on $[a, b]$ and let $f(a) = \alpha$ and $f(b) = \beta$. Then $y = f(x)$ has an inverse function $x = \varphi(y)$ on $[\alpha, \beta]$, $x = \varphi(y)$ being continuous and increasing on $[\alpha, \beta]$.*

We confine ourselves to a geometric interpretation (Fig. 8.11) of Theorem 8.4. The curve AB represents the graph of $y = f(x)$ which is continuous and increasing on $[a, b]$. Every y in $[\alpha, \beta]$ corresponds to the only x in $[a, b]$ so that $f(x) = y$. Thus AB sets a one-to-one correspondence between x and y and we may consider x as a function of y on $[\alpha, \beta]$, i.e., we may say that $x = \varphi(y)$ is the inverse function of $y = f(x)$. The function $x = \varphi(y)$ represented by the curve AB is continuous and increasing on $[\alpha, \beta]$ since as y increases, so does x.

Similar consideration is fully applicable to any continuous function which is decreasing on $[a, b]$.

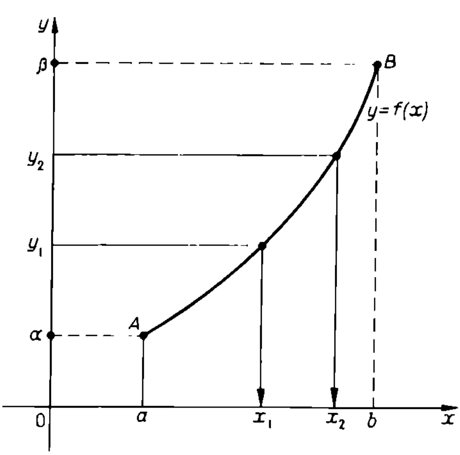


Fig. 8.11

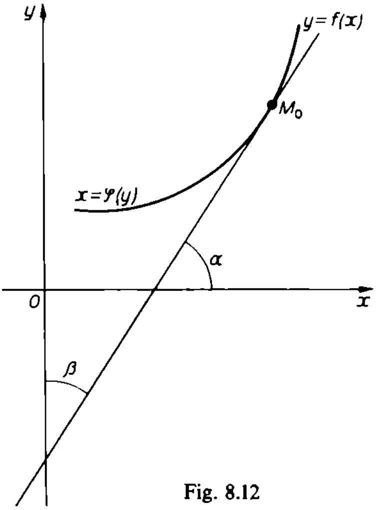


Fig. 8.12

Theorem 8.5. *Let $y = f(x)$ have a derivative $f'(x_0) \neq 0$ at a point x_0 and let there exist an inverse function $x = \varphi(y)$ of $y = f(x)$, $x = \varphi(y)$ being continuous at $y_0 = f(x_0)$. Then the inverse function $x = \varphi(y)$ has a derivative at the point y_0 and*

$$\varphi'(y_0) = \frac{1}{f'(x_0)} \quad (f'(x_0) \neq 0).$$

◄ Consider $x = \varphi(y)$ and an increment Δy in y for $y = y_0$. Then we can write the corresponding increment Δx in $x = \varphi(y)$ as

$$\Delta x = \varphi(y_0 + \Delta y) - \varphi(y_0).$$

Notice that $\Delta y = f(x_0 + \Delta x) - f(x_0)$. Then $\Delta y \neq 0$ implies that $\Delta x \neq 0$; otherwise the given point x_0 would correspond to two distinct values $y_0 = f(x_0)$ and $y_0 + \Delta y$ which is impossible by the definition of a function. Against this background we can write the ratio $\Delta x/\Delta y$ $(\Delta y \neq 0)$ as

$$\frac{\Delta x}{\Delta y} = \frac{1}{\Delta y/\Delta x}. \tag{$*$}$$

If Δy tends to zero so does Δx, for $x = \varphi(y)$ is continuous at y_0.

Since $y = f(x)$ has a derivative at x_0, $\lim\limits_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = f'(x_0)$. We also have $f'(x_0) \neq 0$. Hence there exists a limit of the quotient $\frac{1}{\Delta y/\Delta x}$ as $\Delta y \to 0$ (and, consequently, as $\Delta x \to 0$) and this limit is equal to $\frac{1}{f'(x_0)}$. Then $(*)$ implies that there exists a limit of $\frac{\Delta x}{\Delta y}$ as $\Delta y \to 0$ and

$$\lim_{\Delta y \to 0} \frac{\Delta x}{\Delta y} = \frac{1}{f'(x_0)}.$$

On the other hand, the limit of $\frac{\Delta x}{\Delta y}$ as $\Delta y \to 0$ is the derivative $\varphi'(y_0)$ of $x = \varphi(y)$ at the point $y = y_0$. Hence we have

$$\varphi'(y_0) = \frac{1}{f'(x_0)} \quad \text{or} \quad x'_y = \frac{1}{y'_x}. \quad \blacktriangleright \tag{$**$}$$

It is easy to interpret Theorem 8.5 geometrically. If $y = f(x)$ has a derivative at x_0 then there exists a tangent to a graph of $y = f(x)$ at $M_0(x_0, f(x_0))$ and if this tangent is not parallel to the x-axis, then it is also a tangent to the curve $x = \varphi(y)$ at M_0 (Fig. 8.12). (Notice that the functions $y = f(x)$ and $x = \varphi(y)$ are inverse to each other and are specified by the same curve.) Looking at Fig. 8.12 we see that $f'(x_0) = \tan\alpha$, $\varphi'(y_0) = \tan\beta$ and $\alpha + \beta = \pi/2$ so that $\tan\beta = \tan(\pi/2 - \alpha) = \cot\alpha = 1/\tan\alpha$, i.e.,

$$\varphi'(y_0) = \frac{1}{f'(x_0)}.$$

We can also write $(**)$ as

$$f'(x_0) = \frac{1}{\varphi'(y_0)} \quad \text{or} \quad y'_x = \frac{1}{x'_y} \quad (x'_y \neq 0).$$

Differentiation formulas for inverse functions are easily obtainable by applying the chain rule. Indeed, let $y = f(x)$ and $x = \varphi(y)$ be mutually inverse functions. Then $\varphi[f(x)] = x$. On differentiating both sides with respect to x and using the chain rule, we get

$$\begin{tikzcd} & \varphi'_y = \dfrac{1}{f'_x} \quad (f'_x \neq 0), \\ \varphi'_y f'_x = 1 \arrow[ur, Rightarrow] \arrow[dr, Rightarrow] & \\ & f'_x = \dfrac{1}{\varphi'_y} \quad (\varphi'_y \neq 0). \end{tikzcd}$$

Differentiation of inverse trigonometric functions. (a) The function $y = \sin^{-1} x$ is defined on the closed interval $[-1, 1]$ (Fig. 8.13). Notice that $y = \sin^{-1} x$ is the inverse of $x = \sin y$ defined on $-\pi/2 \leqslant y \leqslant \pi/2$.

On the open interval $-1 < x < 1$ $x = \sin y$ has the positive derivative $x'_y = \cos y$ for each $y \in (-\pi/2, \pi/2)$. Then there exists the derivative

$$y'_x = \frac{1}{x'_y} = \frac{1}{\cos y} = \frac{1}{\sqrt{1 - \sin^2 y}} = \frac{1}{\sqrt{1 - x^2}},$$

where $-1 < x < 1$ and the "+" sign of $\sqrt{1 - x^2}$ is chosen, since $\cos y > 0$ for all $y \in (-\pi/2, \pi/2)$.

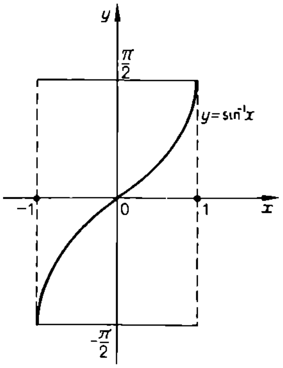


Fig. 8.13

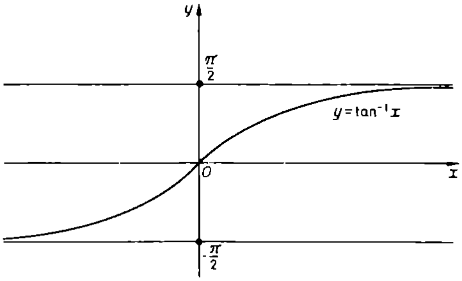


Fig. 8.14

Hence

$$(\sin^{-1} x)' = \frac{1}{\sqrt{1 - x^2}}, \quad -1 < x < 1. \tag{$*$}$$

The points $x = \pm 1$ are deleted, for the derivative $x'_y = \cos y$ is equal to zero at $y = \pm\pi/2$ where the right-hand side of $(*)$ becomes undefined.

(b) The function $y = \tan^{-1} x$, $-\infty < x < +\infty$, (Fig. 8.14), is the inverse function of $x = \tan y$, $-\pi/2 < y < \pi/2$. Then

$$y'_x = \frac{1}{x'_y} = \frac{1}{\dfrac{1}{\cos^2 y}} = \frac{1}{1 + \tan^2 y} = \frac{1}{1 + x^2}.$$

Hence

$$(\tan^{-1} x)' = \frac{1}{1 + x^2} \quad \text{for all } x.$$

(c) To derive the formulas for the derivatives of $\cos^{-1} x$ and $\cot^{-1} x$ it suffices to notice that

$$\sin^{-1} x + \cos^{-1} x = \pi/2,$$
$$\tan^{-1} x + \cot^{-1} x = \pi/2.$$

Whence

$$(\cos^{-1} x)' = -\frac{1}{\sqrt{1 - x^2}} \quad \text{for } -1 < x < 1,$$

$$(\cot^{-1} x)' = -\frac{1}{1 + x^2} \quad \text{for all } x.$$

Differentiation of hyperbolic functions. By definition

$$\sinh x = \frac{e^x - e^{-x}}{2} \quad \text{and} \quad \cosh x = \frac{e^x + e^{-x}}{2}.$$

Then

$$(\sinh x)' = \left(\frac{e^x - e^{-x}}{2}\right)' = \cosh x$$

and

$$(\cosh x)' = \left(\frac{e^x + e^{-x}}{2}\right)' = \sinh x.$$

Also by definition

$$\tanh x = \frac{\sinh x}{\cosh x} \quad \text{and} \quad \coth x = \frac{\cosh x}{\sinh x}.$$

Using the *quotient rule* and the identity $\cosh^2 x - \sinh^2 x \equiv 1$, we obtain

$$(\tanh x)' = \left(\frac{\sinh x}{\cosh x}\right)' = \frac{\cosh^2 x - \sinh^2 x}{\cosh^2 x} = \frac{1}{\cosh^2 x}$$

and

$$(\coth x)' = \left(\frac{\cosh x}{\sinh x}\right)' = \frac{\sinh^2 x - \cosh^2 x}{\sinh^2 x} = -\frac{1}{\sinh^2 x} \quad (x \neq 0).$$

Differentiation of basic elementary functions. Below is the list of formulas widely used in differential calculus.

$$(x^\alpha)' = \alpha x^{\alpha-1} \qquad (\alpha \text{ is any real number, } x > 0);$$

$$(\log_a x)' = \frac{1}{x \ln a} \qquad (a > 0,\ a \neq 1,\ x > 0);$$

$$(\ln x)' = \frac{1}{x} \qquad (x > 0);$$

$$(a^x)' = a^x \ln a \qquad (a > 0,\ a \neq 1);$$

$$(e^x)' = e^x;$$

$$(\sin x)' = \cos x;$$

$$(\cos x)' = -\sin x;$$

$$(\tan x)' = \frac{1}{\cos^2 x} \qquad \left(x \neq \frac{\pi}{2} + n\pi,\ n = 0,\ \pm 1,\ \pm 2,\ \ldots\right);$$

$$(\cot x)' = -\frac{1}{\sin^2 x} \qquad (x \neq n\pi,\ n = 0,\ \pm 1,\ \pm 2,\ \ldots);$$

$$(\sin^{-1} x)' = \frac{1}{\sqrt{1 - x^2}} \qquad (-1 < x < 1);$$

$$(\cos^{-1} x)' = -\frac{1}{\sqrt{1 - x^2}} \qquad (-1 < x < 1);$$

$$(\tan^{-1} x)' = \frac{1}{1 + x^2},$$

$$(\cot^{-1} x)' = -\frac{1}{1 + x^2};$$

$$(\sinh x)' = \cosh x;$$

$$(\cosh x)' = \sinh x;$$

$$(\tanh x)' = \frac{1}{\cosh^2 x};$$

$$(\coth x)' = -\frac{1}{\sinh^2 x} \qquad (x \neq 0);$$

Logarithmic differentiation. Sometimes it is much easier to find a derivative of a natural logarithm of a function than that of this function itself. This gives rise to a convenient differentiating procedure called the *logarithmic differentiation.*

Suppose we are seeking for the derivative of a positive function $y = f(x)$ and assume that it is much easier to compute the derivative of $\varphi(x) = \ln y = \ln f(x)$. Then differentiating this function with respect to x, we obtain

$$\frac{y'}{y} = \varphi'(x).$$

Whence $y' = y \cdot \varphi'(x)$ or

$$y' = f(x)(\ln f(x))'.$$

This procedure is especially helpful when we deal with *composite exponential* functions (*power-exponential* functions) of the form

$$y = [u(x)]^{v(x)},$$

where $u(x)$ and $v(x)$ are both differentiable and $u(x) > 0$.

Then, by logarithmic differentiation we have

$$\ln y = v(x) \ln u(x)$$

and

$$\frac{y'}{y} = v'(x) \ln u(x) + v(x) \frac{u'(x)}{u(x)}.$$

Whence

$$y'(x) = [u(x)]^{v(x)} \left(v'(x) \ln u(x) + v(x) \frac{u'(x)}{u(x)} \right).$$

Example. Find the derivative of $y = x^x$, $x > 0$.

◄ Taking the natural logarithm of $y = x^x$, we have

$$\ln y = x \ln x.$$

Differentiating with respect to x gives

$$\frac{y'}{y} = \ln x + 1.$$

Whence

$$y' = y(\ln x + 1)$$

or

$$y' = x^x(\ln x + 1). \quad ►$$

Applying differentials to approximate computations. Let $y = f(x)$ be differentiable at a point x so that an increment in y admits a representation of the form

$$\Delta y = f'(x)\Delta x + \alpha(\Delta x)\Delta x \quad (\Delta x \neq 0),$$

where $f'(x)\Delta x = dy$ and $\alpha(\Delta x) \to 0$ as $\Delta x \to 0$.

Let $dy \neq 0$ and, consequently, $f'(x) \neq 0$. Then

$$\frac{\Delta y}{dy} = 1 + \frac{\alpha(\Delta x)}{f'(x)}.$$

Whence $\lim_{\Delta x \to 0} \frac{\Delta y}{dy} = 1$, i.e., the infinitesimals Δy and dy are equivalent as $\Delta x \to 0$. This means that the difference $\Delta y - dy = \alpha(\Delta x)\,\Delta x$ is an infinitesimal of higher order than Δy and dy as $\Delta x \to 0$. Hence we may use dy as an approximation of Δy, i.e., $\Delta y \simeq dy$, and the relative error can be made whatever small for any sufficiently small $|\Delta x|$. Then we can apply the relation

$$f(x + \Delta x) \simeq f(x) + f'(x)\Delta x$$

to approximate values of $f(x)$.

Example. Let $y = x^\beta$ (β is any real number). Then

$$\Delta y = (x + \Delta x)^\beta - x^\beta,$$
$$dy = \beta x^{\beta - 1}\Delta x.$$

Given small $|\Delta x|$, we have

$$(x + \Delta x)^\beta \simeq x^\beta + dy$$

or

$$(x + \Delta x)^\beta \simeq x^\beta + \beta x^{\beta - 1}\Delta x.$$

In particular, for $\beta = 1/2$

$$\sqrt{x + \Delta x} \simeq \sqrt{x} + \frac{1}{2\sqrt{x}}\Delta x \quad (x \neq 0).$$

For instance, for $\sqrt{3.9978} = \sqrt{4 + (-0.0022)}$. Putting $x = 4$ and $\Delta x = -0.0022$, we get

$$\sqrt{3.9978} = \sqrt{4 + (-0.0022)} \simeq \sqrt{4} + \frac{1}{2\sqrt{4}}(-0.0022) = 1.99945.$$

8.4 Derivatives and Differentials of Higher Orders

Derivatives of higher orders. Let $f(x)$ have a derivative at every point x in an open interval (a, b). Then the derivative $f'(x)$ of $f(x)$ is a function defined on (a, b). It may happen that $f'(x)$ has a derivative at a point x in (a, b). This derivative of $f'(x)$ is called the *second derivative* of $f(x)$ or the derivative of the second order of $f(x)$ and is denoted by $f''(x)$ or by $f^{(2)}(x)$. Thus

$$f''(x) = (f'(x))'.$$

Similarly, the nth-order derivative of $f(x)$ is the derivative of the $(n-1)$th-order derivative of $f(x)$, i.e.,

$$f^{(n)}(x) = (f^{(n-1)}(x))'.$$

To compute the nth derivative $f^{(n)}(x)$ one has to find the first derivative $f'(x)$ of $f(x)$, then the second derivative $f''(x)$ as the first derivative of $f'(x)$ and continue this process until the nth derivative is computed. Therefore the rules and formulas of computing the first derivatives are sufficient to obtain derivatives of any desired order.

Examples. (1) Compute the nth derivative of $y = e^{kx}$ where $k = \text{const}$.
◄ We get in succession $y' = ke^{kx}$, $y'' = k^2e^{kx}$, $y''' = k^3e^{kx}$, $\ldots$. By induction we can easily prove that

$$(e^{kx})^{(n)} = k^ne^{kx} \quad (n = 1, 2, 3, \ldots). \blacktriangleright$$

(2) Compute the nth derivative of $y = \sin x$.
◄ We have $y' = \cos x = \sin\left(x + \frac{\pi}{2}\right)$, $y'' = -\sin x = \sin(x + \pi) = \sin\left(x + 2\frac{\pi}{2}\right)$, $\ldots$.

By induction it follows that

$$(\sin x)^{(n)} = \sin\left(x + n\frac{\pi}{2}\right) \quad \text{for all } n \in \mathbb{N}. \quad \blacktriangleright$$

(3) Compute the nth derivative of $y = \cos x$.
◄ Following the usual approach we have

$$(\cos x)^{(n)} = \cos\left(x + n\frac{\pi}{2}\right) \quad \text{for all } n \in \mathbb{N}.$$

The set of all functions $f(x)$ which are defined on (a, b) and have continuous nth derivatives at every point $x \in (a, b)$ is denoted by $C^n(a, b)$. We say that a function $f(x)$ is *infinitely differentiable* on (a, b) and write $f(x) \in C^\infty(a, b)$ if $f(x)$ has derivatives of any order at each point $x \in (a, b)$. For example, the functions e^x, $\sin x$ and $\cos x$ are infinitely differentiable on $(-\infty, +\infty)$. ►

(4) Compute all the derivatives of the function $y = x^4$.
◄ We have $y^{(1)} = 4x^3$, $y^{(2)} = 12x^2$, $y^{(3)} = 24x$, $y^{(4)} = 24$. Since $y^{(4)}$ is a constant all derivatives of higher orders are zero, i.e., $y^{(5)} = y^{(6)} = \ldots = y^{(n)} = 0$. ►

To give a physical interpretation of the second derivative we consider the law $s = s(t)$ of a rectilinear motion of a point. Then the first derivative $s'(t) = v(t)$ specifies the velocity of the point at the moment t and the second derivative $s''(t) = v'(t)$ is the acceleration of the point at t.

Leibniz formula. Let $u(x)$ and $v(x)$ have the nth derivatives. Then

(a) if $y(x) = u(x) \pm v(x)$ then $y^{(n)}(x) = u^{(n)}(x) \pm v^{(n)}(x)$;

(b) if $y(x) = u(x)v(x)$ then

$$\begin{aligned} y' &= u'v + uv', \\ y'' &= u''v + 2u'v' + uv'', \\ y''' &= u'''v + 3u''v' + 3u'v'' + uv''' \quad \text{and so on.} \end{aligned}$$

It is easy to notice that the formulas on the right resemble the binomial formulas for $(u + v)^1$, $(u + v)^2$ and $(u + v)^3$ with the exponents indicating how many times u and v are to be differentiated. The resemblance becomes nearly complete if u and v are replaced by $u^{(0)}$ and $v^{(0)}$, standing for derivatives of the zero order. Using the method of mathematical induction, we can show that in general

$$(uv)^{(n)} = n^{(n)} \cdot v + \frac{n}{1!} u^{(n-1)} v^{(1)} + \frac{n(n-1)}{2!} u^{(n-2)} v^{(2)}$$

$$+ \dots + \frac{n(n-1) \dots (n-k+1)}{k!} u^{(n-k)} \cdot v^{(k)} + \dots + uv^{(n)}.$$

This relation is called the *Leibniz formula*.

Example. Using the Leibniz formula, compute the derivative $y^{(1001)}$ of the function $y = x^2 e^x$.

◄ We have

$$y = e^x \cdot x^2 = (e^x)^{(1001)} x^2 + \frac{1001}{1!} (e^x)^{(1000)} \cdot (x^2)'$$

$$+ \frac{1001 \times 1000}{2!} (e^x)^{(999)} (x^2)'' + 0 = e^x x^2 + 2002 e^x x + 1001 \times 10^3 e^x. \quad ►$$

We shall derive another useful formula. Let $x = \varphi(y)$ and $y = f(x)$ be mutually inverse functions and let $f'(x) \neq 0$. Then

$$x'_y = \frac{1}{y'_x}$$

and

$$x''_y = \frac{d}{dy} (x'_y) = \frac{d}{dy} \left(\frac{1}{y'_x} \right) = \frac{d}{dx} \left(\frac{1}{y'_x} \right) \frac{dx}{dy} = - \frac{y''_x}{(y'_x)^2} \frac{1}{y'_x} = - \frac{y''_x}{(y'_x)^3}.$$

Differentials of higher orders. Let $y = f(x)$ be differentiable at a point x. It may happen that at x the differential $dy = f'(x)\,dx$ is a differentiable function of x. Then there exists a differential of a differential of a given function which is called the *differential of the second order* or the *second* *differential* of $y = f(x)$ and is denoted by d^2y so that

$$d^2y = d(dy).$$

Analogously we can define differentials of higher orders. The *differential of the nth order* d^ny or the *nth differential* of a function $y = f(x)$ is the differential of the $(n - 1)$th differential of $y = f(x)$ so that

$$d^ny = d(d^{n-1}y).$$

The differential dy may be called the *first differential* or the *differential of the first order* of a function $y = f(x)$.

Now we shall derive some important formulas for differentials of higher orders. Let $y = f(x)$ be a function of an independent variable x and let $y = f(x)$ have differentials of any order. Then

$$dy = f'(x)\,dx,$$

where $dx = \Delta x$ is independent of x.

By definition

$$d^2y = d(dy) = d(f'(x)\,dx),$$

where $f'(x)dx$ is a function of x, so that the factor dx being independent of x can be taken outside the differential sign. So

$$d^2y = d(f'(x))dx.$$

Computing $d(f'(x))$ as the first differential of $f'(x)$, we obtain

$$d(f'(x)) = (f'(x))'\,dx = f''(x)\,dx.$$

Hence the second-order differential d^2y of the function $y = f(x)$ at the point x is given by the formula

$$d^2y = f''(x)dx^2,$$

where dx^2 stands for $(dx)^2$, dx being the differential of the independent variable x at the point x.

Mathematical induction gives the formula of the nth differential as

$$d^ny = f^{(n)}(x)dx^n,$$

where $dx^n = (dx)^n$.

Whence

$$f^{(n)}(x) = \frac{d^ny}{dx^n}.$$

Now let $y = f(u)$ where $u = \varphi(x)$ is a function differentiable sufficiently many times. Since the first differential retains its form, we have

$$dy = f'(u)du,$$

where $du = \varphi'(x)dx$ is in general dependent on x. Then

$$\begin{aligned} d^2y &= d(dy) = d(f'(u)du) = d(f'(u))du + f'(u)d(du) \\ &= f''(u)du^2 + f'(u)d^2u. \end{aligned}$$

If u is an independent variable then $d^2u = 0$ and

$$d^2y = f''(u)du^2.$$

Comparing both expressions for d^2y, we conclude that the second differential does not possess the property of invariance of the form. However, given a linear function $u = \varphi(x) = ax + b$ where a and b are constant, the second differential retains its form.

Differentiation of functions given in a parametric form. Suppose that the Cartesian coordinate xy-system is set up on a plane. Let $\varphi(t)$ and $\psi(t)$ be continuous functions on a closed interval $\alpha \leqslant t \leqslant \beta$. If we regard the parameter t as time the functions $\varphi(t)$ and $\psi(t)$ specify the law of motion of a point M in the xy-plane so that the coordinates of M are given by

$$\begin{cases} x = \varphi(t) \\ y = \psi(t) \end{cases} \quad \alpha \leqslant t \leqslant \beta. \tag{*}$$

The set M of all points whose coordinates (x, y) on the xy-plane are given by the equations (*) is called the *plane curve*. In this case we say that a plane curve is given in a *parametric form* or is represented parametrically.

For example, we can parametrically represent a circle with radius R and centre at the origin of coordinates by the equations

$$\begin{cases} x = R\cos t \\ y = R\sin t \end{cases} \quad 0 \leqslant t < 2\pi,$$

where t is the angle between the x-axis and the position-vector $\overrightarrow{OM}$ from the origin O to the point $M(x, y)$ measured in radians (Fig. 8.15).

The parametric equations of a plane curve can be reduced to the equation $F(x, y) = 0$ by eliminating the parameter t. For example, if we square and add the parametric equations of the circle the parameter t is eliminated so that the circle with radius R and centre at the origin of coordinates is given by the equation $x^2 + y^2 = R^2$, familiar to the reader. However, sometimes we fail to eliminate t from parametric equations. In these cases we need techniques and procedures that enable us to compute the derivative of y with respect to x when the curve is given in a parametric form.

We say that a functional relationship between y and x is represented *parametrically* if both variables x and y are specified separately as functions of a parameter t so that $x = \varphi(t)$, $y = \psi(t)$, $t \in (\alpha, \beta)$.

Let $x = \varphi(t)$ and $y = \psi(t)$ be defined and continuous for the values of

t in the open interval (α, β). Suppose that there exists an inverse function $t = g(x)$ of $x = \varphi(t)$. Then $y = \psi[g(x)]$ is a composite function of x. Furthermore, we assume that $\varphi(t)$ and $\psi(t)$ are differentiable at the point $t \in (\alpha, \beta)$, $\varphi'(t) \neq 0$ and $t = g(x)$ is differentiable at a point x corresponding to t. Then by Theorem 8.3 the function $y = \psi[g(x)]$ is differentiable at the point x and

$$y'_x = y'_t t'_x.$$

On the other hand, by virtue of Theorem 8.5 we have

$$t'_x = \frac{1}{x'_t}$$

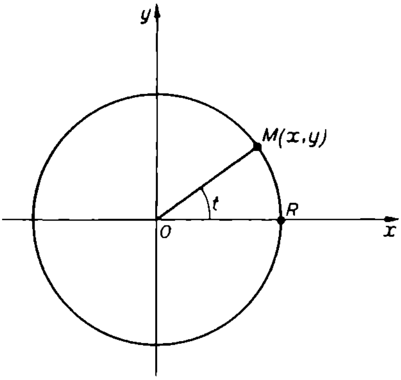


Fig. 8.15

so that

$$y'_x = y'_t \frac{1}{x'_t} = \frac{y'_t}{x'_t}$$

or

$$\frac{dy}{dx} = \frac{\psi'(t)}{\varphi'(t)} \quad (\varphi'(t) \neq 0).$$

This result immediately follows when we divide the nominator and denominator of $\frac{dy}{dx}$ by dt so that

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{\psi'(t)}{\varphi'(t)}.$$

Example. Consider the circle represented parametrically by

$$\begin{cases} x = R\cos t \\ y = R\sin t \end{cases} \quad 0 \leqslant t < 2\pi.$$

Then

◄ $$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{R\cos t}{-R\sin t} = -\cot t$$

or $\dfrac{dy}{dx} = -\dfrac{x}{y}$ (interpret this result geometrically). ►

If $\varphi(t)$ and $\psi(t)$ have the kth derivatives and $\varphi'(t) \neq 0$ then the composite function $y = \psi[g(x)]$ has the kth derivative with respect to x.

The second derivative of y with respect to x is given by

$$\frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{d}{dt}\left(\frac{\psi'(t)}{\varphi'(t)}\right)\frac{dt}{dx}$$

$$= \frac{\psi''(t)\varphi'(t) - \psi'(t)\varphi''(t)}{(\varphi')^2(t)} \cdot \frac{1}{\varphi'(t)} = \frac{\psi''(t)\varphi'(t) - \psi'(t)\varphi''(t)}{(\varphi')^3(t)}.$$

Hence

$$y_x'' = \frac{(y_x)_t'}{x_t'}$$

and in general

$$y_x^{(n)} = \frac{[y_x^{(n-1)}]_t'}{x_t'},$$

where $y = f(x)$ is given in a parametric form by the equations $x = x(t)$ and $y = y(t)$.

Example. Compute $\dfrac{d^2y}{dx^2}$ provided that $y = f(x)$ is given by

$$\begin{cases} x = a(t - \sin t), \\ y = a(1 - \cos t). \end{cases}$$

◄ We have

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{a\sin t}{a(1 - \cos t)} = \cot\frac{t}{2}$$

and

$$\frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{d}{dt}\left(\cot\frac{t}{2}\right)\frac{dt}{dx} = \frac{d}{dt}\left(\cot\frac{t}{2}\right)\frac{1}{dx/dt}$$

$$= -\frac{1}{\sin^2\frac{t}{2}} \cdot \frac{1}{2} \cdot \frac{1}{a(1 - \cos t)} = -\frac{1}{4a\sin^4\frac{t}{2}}. \quad ►$$

8.5 Mean Value Theorems

Theorem 8.6 (Rolle's theorem). *If a function* $f(x)$ (1) *is continuous on* $[a, b]$, (2) *has a finite derivative at every point of* (a, b) *and* (3) *assumes equal values at the endpoints of* $[a, b]$ *so that* $f(a) = f(b)$, *then there exists at least one point* $\xi \in (a, b)$ *such that* $f'(\xi) = 0$.

◀ Since $f(x)$ is continuous on $[a, b]$ Theorem 7.30 implies that $f(x)$ attains its maximum M and minimum m on $[a, b]$.

Two cases should be distinguished:

(a) Let $M = m$. Then $M \leqslant f(x) \leqslant M$, i.e., $f(x)$ is a constant function on $[a, b]$. Hence $f'(x) = 0$ for all x in (a, b) and the theorem is true.

(b) Let $M \neq m$. Then $f(x)$ attains either a maximum M or a minimum m at a point ξ contained in (a, b) since $f(a) = f(b)$ so that $f(x)$ cannot attain its maximum M at one end of $[a, b]$ and its minimum m at the other end simultaneously. For definiteness we set $M = f(\xi)$, $a < \xi < b$ (Fig. 8.16).

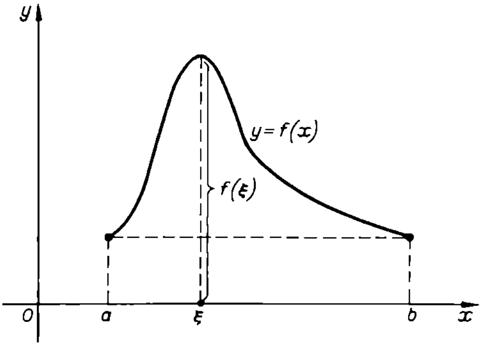


Fig. 8.16

Since $f(x)$ has a derivative $f'(x)$ at every point of (a, b) there exists a derivative $f'(\xi)$ at ξ and

$$\lim_{\substack{\Delta x \to 0 \\ \Delta x > 0}} \frac{f(\xi + \Delta x) - f(\xi)}{\Delta x} = \lim_{\substack{\Delta x \to 0 \\ \Delta x > 0}} \frac{f(\xi - \Delta x) - f(\xi)}{-\Delta x} = f'(\xi).$$

On the other hand, $f(\xi) = M$ is the maximum of $f(x)$ on $[a, b]$ so that

$$f(\xi + \Delta x) - f(\xi) \leqslant 0 \quad \text{and} \quad f(\xi - \Delta x) - f(\xi) \leqslant 0.$$

Whence

$$\frac{f(\xi + \Delta x) - f(\xi)}{\Delta x} \leqslant 0 \quad \text{and} \quad \frac{f(\xi - \Delta x) - f(\xi)}{-\Delta x} \geqslant 0 \quad (\Delta x > 0).$$

Evaluating the limits as $\Delta x \to 0$, we obtain the inequalities

$$f'(\xi) \leqslant 0 \quad \text{and} \quad f'(\xi) \geqslant 0$$

which must be simultaneously true. Hence, $f'(\xi) = 0$ and the theorem is proved. ►

The geometric interpretation of Rolle's theorem is that if $y = f(x)$ satisfies all the three conditions of the theorem then the graph of $f(x)$ is represented by the curve AB (Fig. 8.17) such that (1) the curve AB is continuous on $[a, b]$, (2) there exists a tangent to the curve at any point between $A(a, f(a))$ and $B(b, f(b))$ and (3) the ordinates of the points A and B of this curve are equal. Rolle's theorem states that at least for one point C $(\xi, f(\xi))$ of the curve AB there is a tangent parallel to the x-axis.

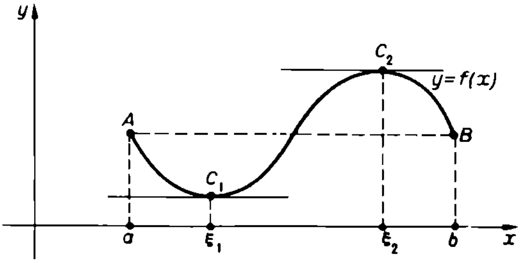


Fig. 8.17

By way of illustration we shall show that the conditions (1)-(3) are important and if they are violated Rolle's theorem may not be true. Consider the function $f(x) = |x|$, $-1 \leqslant x \leqslant 1$ (Fig. 8.18). For this function condition (2) is not satisfied as $f(x)$ has no derivative at $x = 0$. In this case Rolle's theorem is inapplicable as in $(-1, 1)$ there exists no point at which the derivative $f'(x)$ is equal to zero. Indeed, $f'(x) = -1$ if $-1 < x < 0$ and $f'(x) = 1$ if $0 < x < 1$ while at $x = 0$ $f'(x)$ does not exist. Looking at Fig. 8.19 we notice that the function $f(x) = x - [x]$ does not satisfy condi-

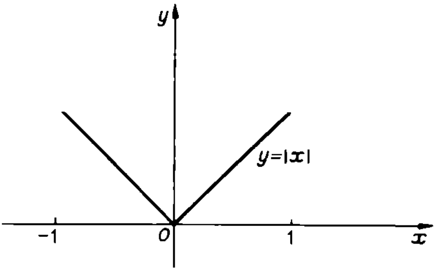


Fig. 8.18

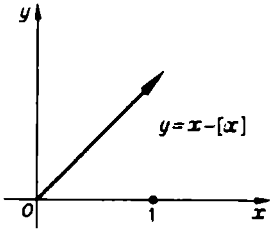


Fig. 8.19

tion (1) of Rolle's theorem since the point $x = 1$ is a discontinuity of $f(x)$ and thus $f(x)$ is not continuous on $[0, 1]$; the derivative $f'(x)$ is equal to 1 at all points in $(0, 1)$.

Problems. (1) Consider the function $f(x) = 1 + x^m(1 - x)^n$ where m and n are positive integers. Without computing the derivative of $f(x)$, show that the equation $f'(x) = 0$ has at least one real root in $(0, 1)$.

(2) Prove that the equation $x^3 + 3x - 6 = 0$ has only one real root.

Theorem 8.7 (mean value theorem)[*]. *If a function* $f(x)$ (1) *is continuous on* $[a, b]$ *and* (2) *has a derivative* $f'(x)$ *in* (a, b) *then there exists at least one point* ξ *in* (a, b) *such that*

$$\frac{f(b) - f(a)}{b - a} = f'(\xi), \quad a < \xi < b.$$

◄ We introduce an auxiliary function $F(x)$ on $[a, b]$ as

$$F(x) = f(x) - f(a) - \frac{f(b) - f(a)}{b - a}(x - a).$$

The function $F(x)$ satisfies the conditions of Rolle's theorem. Indeed, $F(x)$ is continuous on $[a, b]$ since every summand in $F(x)$ is continuous on $[a, b]$; in (a, b) $F(x)$ has a finite derivative, for every summand of $F(x)$ has a derivative in (a, b); and $F(a) = F(b) = 0$, i.e., $F(x)$ assumes equal values at the endpoints of $[a, b]$.

Then by virtue of Rolle's theorem there exists at least one point ξ in (a, b) such that $F'(\xi) = 0$.

On the other hand

$$F'(x) = f'(x) - \frac{f(b) - f(a)}{b - a}$$

so that at the point ξ

$$f'(\xi) - \frac{f(b) - f(a)}{b - a} = 0.$$

Whence

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}, \quad \xi \in (a, b). \ ►$$

Rolle's theorem is a specific case of the mean value theorem for derivatives and is easily obtained by putting $f(a) = f(b)$ in the latter.

---

[*] In Russian-language mathematical literature this theorem is called the *Lagrange theorem*.

To interpret the mean value theorem geometrically we turn to Fig. 8.20. It is easy to notice that the ratio $\frac{f(b) - f(a)}{b - a}$ is the slope of the line AB and $f'(\xi)$ is the slope of the tangent to $y = f(x)$ at the point with the abscissa $x = \xi$. Hence the mean value theorem states that if AB is a continuous curve such that there is a tangent at every point between A and B then there exists at least one point $C(\xi, f(\xi))$ between A and B at which the tangent to the curve AB is parallel to the straight line AB.

The formula

$$\frac{f(b) - f(a)}{b - a} = f'(\xi)$$

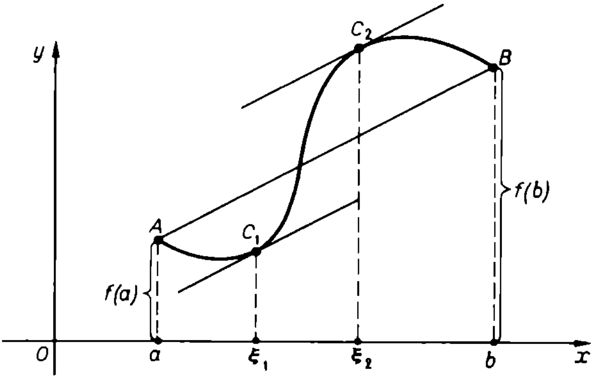


Fig. 8.20

or

$$f(b) - f(a) = f'(\xi)(b - a), \quad a < \xi < b$$

is also valid for $a > b$. Since ξ is in general unknown it is convenient to represent it as

$$\xi = a + \theta(b - a),$$

where θ is a real number, $0 < \theta < 1$. Then the above formula becomes

$$f(b) - f(a) = f'(a + \theta(b - a))(b - a), \quad 0 < \theta < 1.$$

Replacing a and b by x and $x + \Delta x$, respectively, we get

$$\Delta f(x) = f(x + \Delta x) - f(x) = f'(x + \theta \Delta x)\Delta x, \quad 0 < \theta < 1.$$

This is the exact expression that relates an increment of $y = f(x)$ and an increment Δx whereas the relative error of the approximate relation

$$\Delta f(x) \simeq f'(x)\Delta x$$

tends to zero only as $\Delta x \to 0$. Also notice that in the exact relation the number θ is in general unknown.

Example. By use of the mean value theorem prove that

$$|\tan^{-1} x_2 - \tan^{-1} x_1| \leqslant |x_2 - x_1| \quad \forall x_1, x_2.$$

◄ Consider the function $f(x) = \tan^{-1} x$. This function satisfies the conditions of the mean value theorem on any $[a, b]$. Then for any x_1 and x_2

$$f(x_2) - f(x_1) = f'(\xi)(x_2 - x_1)$$

or

$$\tan^{-1} x_2 - \tan^{-1} x_1 = \frac{1}{1 + \xi^2}(x_2 - x_1),$$

where the point ξ is between the points x_1 and x_2.

Whence

$$|\tan^{-1} x_2 - \tan^{-1} x_1| = \frac{1}{1 + \xi^2}|x_2 - x_1|$$

and

$$|\tan^{-1} x_2 - \tan^{-1} x_1| \leqslant |x_2 - x_1|$$

since $\dfrac{1}{1 + \xi^2} \leqslant 1$ for all ξ. ►

Problem. Using the mean value theorem prove that

$$\frac{x}{1 + x} < \ln(1 + x) < x, \quad x > -1.$$

Theorem 8.8 (Cauchy mean value theorem). *If the functions $f(x)$ and $\varphi(x)$ (1) are continuous on $[a, b]$, (2) have derivatives $f'(x)$ and $\varphi'(x)$ in (a, b) and (3) if $\varphi'(x) \neq 0$ in (a, b) then there exists at least one point ξ in (a, b) such that*

$$\frac{f(b) - f(a)}{\varphi(b) - \varphi(a)} = \frac{f'(\xi)}{\varphi'(\xi)}, \quad a < \xi < b. \tag{*}$$

◄ From the conditions of the theorem it follows that the difference $\varphi(b) - \varphi(a)$ can not be equal to zero. Indeed, if $\varphi(b) - \varphi(a) = 0$ then $\varphi(x)$ would satisfy the conditions of Rolle's theorem so that $\varphi'(x)$ would be equal to zero at least at one point ξ in (a, b). However this contradicts condition (3) of this theorem. Hence the formula (*) makes sense. Let us show that this formula holds for some ξ in (a, b).

Consider the auxiliary function

$$F(x) = f(x) - f(a) - \frac{f(b) - f(a)}{\varphi(b) - \varphi(a)}(\varphi(x) - \varphi(a))$$

that satisfies the conditions of Rolle's theorem. Indeed, (1) $F(x)$ is continuous on $[a, b]$ since $f(x)$ and $\varphi(x)$ are continuous on $[a, b]$; (2) $F(x)$ has a derivative $F'(x)$ at every point in (a, b), since every summand in the right-hand side of $F(x)$ has a derivative in (a, b); and (3) $F(a) = F(b) = 0$.

From Rolle's theorem we conclude that between a and b there exists ξ such that $F'(\xi) = 0$. The derivative $F'(x)$ of $F(x)$ is

$$F'(x) = f'(x) - \frac{f(b) - f(a)}{\varphi(b) - \varphi(a)} \varphi'(x)$$

so that

$$f'(\xi) - \frac{f(b) - f(a)}{\varphi(b) - \varphi(a)} \varphi'(\xi) = 0.$$

Dividing both sides by $\varphi'(\xi) \neq 0$, we arrive at the desired formula

$$\frac{f'(\xi)}{\varphi'(\xi)} = \frac{f(b) - f(a)}{\varphi(b) - \varphi(a)}. \quad \blacktriangleright$$

It is easy to notice that if we put $\varphi(x) \equiv x$ the Cauchy mean value theorem becomes the mean value theorem for derivatives.

Problem. Given the differences $f(b) - f(a)$ and $\varphi(b) - \varphi(a)$, could we derive the formula involved in the Cauchy mean value theorem by applying the mean value theorem to these differences?

Remark. Rolle's theorem, the mean value theorem and the Cauchy mean value theorem imply that there exists some "middle" point $\xi \in (a, b)$ at which some of the named relations are true. For this reason all these theorems are collectively named the ***mean value theorems for derivatives.***

8.6 L'Hospital's Rule

Let functions $f(x)$ and $\varphi(x)$ be defined in a neighbourhood of a point $x = a$ and let $f(a) = 0$ and $\varphi(a) = 0$. Then the quotient $\frac{f(x)}{\varphi(x)}$ becomes indeterminate at $x = a$ while the limit of this quotient can exist at $x = a$. The notation $\frac{0}{0}$ is frequently used to refer to this ambiguous situation. When we wish to compute $\lim\limits_{x \to a} \frac{f(x)}{\varphi(x)}$ where the quotient $\frac{f(x)}{\varphi(x)}$ is indeterminate at $x = a$ we shall speak of evaluating the ***indeterminate form*** $\frac{0}{0}$.

Evaluating the indeterminate form $\frac{\infty}{\infty}$ involves the computation of $\lim\limits_{x \to a} \frac{f(x)}{\varphi(x)}$ provided that $\lim\limits_{x \to a} f(x) = \infty$ and $\lim\limits_{x \to a} \varphi(x) = \infty$.

Evaluating the indeterminate form $\infty - \infty$ involves the computation of $\lim\limits_{x\to a}[f(x) - \varphi(x)]$ provided that $\lim\limits_{x\to a} f(x) = \infty$ and $\lim\limits_{x\to a}\varphi(x) = \infty$.

We can introduce the similar notions for limits at infinity, i.e., for limits as $x \to \infty$.

Theorem 8.9 (L'Hospital's rule). *Let $f(x)$ and $\varphi(x)$ have derivatives $f'(x)$ and $\varphi'(x)$ in a neighbourhood $(a - \delta, a + \delta)$ of a point a except probably at a. Suppose that $\varphi(x)$ and $\varphi'(x)$ are not equal to zero in $(a - \delta, a + \delta)$. If $\lim\limits_{x\to a} f(x) = 0$ and $\lim\limits_{x\to a}\varphi(x) = 0$ and if the quotient $\dfrac{f'(x)}{\varphi'(x)}$ has a finite or infinite limit as $x \to a$ then there exists*

$$\lim_{x\to a}\frac{f(x)}{\varphi(x)} = \lim_{x\to a}\frac{f'(x)}{\varphi'(x)}.$$

◄ Theorem 8.9 says nothing about the values of $f(x)$ and $\varphi(x)$ at $x = a$. We put $f(a) = 0$ and $\varphi(a) = 0$. Then $\lim\limits_{x\to a} f(x) = f(a)$ and $\lim\limits_{x\to a}\varphi(x) = \varphi(a)$ and the functions $f(x)$ and $\varphi(x)$ become continuous at a so that on $[a, x]$ (or on $[x, a]$) where x is a point in $(a - \delta, a + \delta)$, $f(x)$ and $\varphi(x)$ satisfy the conditions of the Cauchy mean value theorem. Hence, there exists at least one point $\xi = \xi(x)$ between a and x such that

$$\frac{f(x)}{\varphi(x)} = \frac{f(x) - f(a)}{\varphi(x) - \varphi(a)} = \frac{f'(\xi)}{\varphi'(\xi)}. \tag{$*$}$$

If for some x there exist more than one such ξ we choose any of them.

The point ξ is dependent of x and $\xi \to a$ as $x \to a$. As stated, the quotient $\dfrac{f'(x)}{\varphi'(x)}$ has a finite or infinite limit as $x \to a$. This limit is independent of how x tends to a. So the quotient $\dfrac{f'(\xi)}{\varphi'(\xi)}$ has a limit equal to a limit of $\dfrac{f'(x)}{\varphi'(x)}$ as $x \to a$ and, consequently, as $\xi \to a$, so that

$$\lim_{\substack{x\to a\\(\xi\to a)}}\frac{f'(\xi)}{\varphi'(\xi)} = \lim_{x\to a}\frac{f'(x)}{\varphi'(x)}. \tag{$**$}$$

From the identities $(*)$ and $(**)$ we have obtained it follows that

$$\lim_{x\to a}\frac{f(x)}{\varphi(x)} = \lim_{x\to a}\frac{f'(x)}{\varphi'(x)}.$$ ►

The above identity represents L'Hospital's rule which allows, under some specific conditions, to replace a limit of a quotient of functions by a limit of a quotient of their derivatives which is sometimes easier to compute.

Example.

$$\lim_{x\to 0}\frac{1-\cos x}{x^2}=\lim_{x\to 0}\frac{(1-\cos x)'}{(x^2)'}=\lim_{x\to 0}\frac{\sin x}{2x}=\frac{1}{2}.$$

Remarks. (*1*) If the conditions of this theorem are satisfied on the open interval $(a-\delta, a)$ or on $(a, a+\delta)$ L'Hospital's rule is applicable to computation of the limit of $\frac{f(x)}{\varphi(x)}$ as $x\to a-0$ or as $x\to a+0$, respectively.

(*2*) It may happen than the limit of a quotient of derivatives does not exist while the limit of the quotient of the respective functions does exist. As an illustration we consider the functions $f(x)=x^2\sin\frac{1}{x}$ and $\varphi(x)=x$. For the quotient of these functions at $x=0$ we have

$$\lim_{x\to 0}\frac{f(x)}{\varphi(x)}=\lim_{x\to 0}\frac{x^2\sin\frac{1}{x}}{x}=\lim_{x\to 0}x\sin\frac{1}{x}=0.$$

On the other hand, the quotient of the derivatives $\frac{f'(x)}{\varphi'(x)}=2x\sin\frac{1}{x}-\cos\frac{1}{x}$ has no limit at $x=0$. Hence, from the existence of $\lim\limits_{x\to 0}\frac{f(x)}{\varphi(x)}$ it does not necessarily follow that $\lim\limits_{x\to 0}\frac{f'(x)}{\varphi'(x)}$ exists.

(*3*) Sometimes we have to apply L'Hospital's rule repeatedly when computing $\lim\limits_{x\to a}\frac{f(x)}{\varphi(x)}$. For example, if $f(x)$ and $\varphi(x)$ and their derivatives $f'(x)$ and $\varphi'(x)$ all satisfy the hypothesis of L'Hospital's rule we can apply the rule to compute $\lim\limits_{x\to a}\frac{f'(x)}{\varphi'(x)}$ and so on.

Example.

$$\lim_{x\to 0}\frac{x-\sin x}{x^3}=\lim_{x\to 0}\frac{(x-\sin x)'}{(x^3)'}=\lim_{x\to 0}\frac{1-\cos x}{3x^2}=\lim_{x\to 0}\frac{\sin x}{6x}=\frac{1}{6}.$$

Theorem 8.10. $\left(\textbf{L'Hospital's rule for an indeterminate form } \frac{\infty}{\infty}\right)$. *Let the functions $f(x)$ and $\varphi(x)$ have the derivatives $f'(x)$ and $\varphi'(x)$ in a neighbourhood $(a-\delta, a+\delta)$ of a point a except probably at a. Suppose that $\varphi(x)$ and $\varphi'(x)$ are not equal to zero in $(a-\delta, a+\delta)$. If $\lim\limits_{x\to a}f(x)=\infty$ and*

$\lim\limits_{x\to a}\varphi(x) = \infty$ *and if* $\dfrac{f'(x)}{\varphi'(x)}$ *has a finite or infinite limit as* $x \to a$ *then there exists*

$$\lim_{x\to a}\frac{f(x)}{\varphi(x)} = \lim_{x\to a}\frac{f'(x)}{\varphi'(x)}\,.$$

Here we can also consider one-sided limits as $x \to a - 0$ or as $x \to a + 0$ (see *Remark 1*).

Example.

$$\lim_{x\to 0+0}\frac{\ln\sin ax}{\ln\sin bx} = \lim_{x\to 0+0}\frac{(\ln\sin ax)'}{(\ln\sin bx)'} = \lim_{x\to 0+0}\frac{\dfrac{a\cos ax}{\sin ax}}{\dfrac{b\cos bx}{\sin bx}}$$

$$= \frac{a}{b}\lim_{x\to 0+0}\left(\frac{\cos ax}{\cos bx}\cdot\frac{\sin bx}{\sin ax}\right) = 1 \quad (a > 0,\ b > 0).$$

L'Hospital's rules are used to compute the following limits:

(a) $\lim\limits_{x\to a}[f(x)\varphi(x)]$ for $\lim\limits_{x\to a} f(x) = 0$ and $\lim\limits_{x\to a}\varphi(x) = \infty$.

◄ It suffices to represent $f(x)\varphi(x)$ as

$$f(x)\varphi(x) = \frac{f(x)}{1/\varphi(x)} \quad \text{or} \quad f(x)\varphi(x) = \frac{\varphi(x)}{1/f(x)}\,.$$

Then the functions on the right satisfy the hypothesis of L'Hospital's rule. ►

Example.

$$\lim_{x\to 0+0}(x\ln x) = \lim_{x\to 0+0}\frac{\ln x}{1/x} = \lim_{x\to 0+0}\frac{1/x}{-1/x^2} = 0.$$

(b) $\lim\limits_{x\to a}[f(x) - \varphi(x)]$ for $\lim\limits_{x\to a} f(x) = \infty$ and $\lim\limits_{x\to a}\varphi(x) = \infty$.

◄ To apply L'Hospital's rule it suffices to write $f(x) - \varphi(x)$ as

$$f(x) - \varphi(x) = \frac{1}{1/f(x)} - \frac{1}{1/\varphi(x)} = \frac{1/\varphi(x) - 1/f(x)}{1/f(x)\cdot 1/\varphi(x)}\,. \quad ►$$

Example.

$$\lim_{x\to 1}\left(\frac{x}{x-1} - \frac{1}{\ln x}\right) = \lim_{x\to 1}\frac{x\ln x - x + 1}{(x-1)\ln x}$$

$$= \lim_{x\to 1}\frac{\ln x}{\ln x + 1 - 1/x} = \lim_{x\to 1}\frac{1/x}{1/x + 1/x^2} = \frac{1}{2}\,.$$

(c) $\lim\limits_{x\to a}[f(x)]^{\varphi x}$ for each of the following cases

(i) $\lim\limits_{x\to a} f(x) = 0$ and $\lim\limits_{x\to a} \varphi(x) = 0$ (0^0);

(ii) $\lim\limits_{x\to a} f(x) = 1$ and $\lim\limits_{x\to a} \varphi(x) = \infty$ (1^∞);

(iii) $\lim\limits_{x\to a} f(x) = +\infty$ and $\lim\limits_{x\to a} \varphi(x) = 0$ (∞^0).

◄ Put

$$y = [f(x)]^{\varphi(x)}.$$

Let us consider

$$\ln y = \varphi(x) \ln f(x)$$

and evaluate

$$\lim_{x\to a} \ln y = \lim_{x\to a} [\varphi(x) \ln f(x)].$$

It is easy to notice that we have to evaluate a limit considered in (a) for each of the cases (i)-(iii).

Suppose that $\lim\limits_{x\to a} \ln y = A$. Then $\lim\limits_{x\to a} y = e^A$, i.e.,

$$\lim_{x\to a} [f(x)]^{\varphi(x)} = e^A.$$ ►

Example. Find $\lim\limits_{x\to 0+0} x^x$.

◄ Put $y = x^x$. Then $\ln y = x \ln x$. Whence

$$\lim_{x\to 0+0} \ln y = \lim_{x\to 0+0} \frac{\ln x}{1/x} = \lim_{x\to 0+0} \frac{1/x}{-1/x^2} = 0,$$

so that

$$\lim_{x\to 0+0} y = e^0 = 1.$$ ►

Theorem 8.11. *Suppose that* (1) *functions* $f(x)$ *and* $\varphi(x)$ *are defined for all* x *such that* $|x|$ *is sufficiently large*; (2) $\lim\limits_{x\to\infty} f(x) = \lim\limits_{x\to\infty} \varphi(x) = 0$ *or* $\lim\limits_{x\to\infty} f(x) = \infty$ *and* $\lim\limits_{x\to\infty} \varphi(x) = \infty$; (3) *there exist derivatives* $f'(x)$ *and* $\varphi'(x)$ $(\varphi'(x) \neq 0)$ *for all* x *such that* $|x|$ *is sufficiently large*; (4) *there exists a finite or infinite limit of the quotient* $\dfrac{f'(x)}{\varphi'(x)}$ *as* $x \to \infty$. *Then*

$$\lim_{x\to\infty} \frac{f(x)}{\varphi(x)} = \lim_{x\to\infty} \frac{f'(x)}{\varphi'(x)}.$$

To verify that Theorem 8.11 is true it suffices to put $x = 1/t$ and use the results of Theorem 8.9 and Theorem 8.10.

Example.

$$\lim_{x \to +\infty} \frac{x^2}{e^x} = \lim_{x \to +\infty} \frac{2x}{e^x} = \lim_{x \to +\infty} \frac{2}{e^x} = 0.$$

Applying L'Hospital's rule repeatedly n times, we obtain

$$\lim_{x \to +\infty} \frac{x^n}{e^x} = \lim_{x \to +\infty} \frac{nx^{n-1}}{e^x} = \ldots = \lim_{x \to +\infty} \frac{n!}{e^x} = 0$$

so that the function e^x increases faster than any power function of x as $x \to +\infty$.

The following example shows that L'Hospital's rule being applicable in evaluating $\lim\limits_{x \to +\infty} \frac{\sqrt{1 + x^2}}{x}$ turns out to be unsuitable for practical purposes. Indeed, applying L'Hospital's rule, we have

$$\lim_{x \to +\infty} \frac{\sqrt{1 + x^2}}{x} = \lim_{x \to +\infty} \frac{\frac{x}{\sqrt{1 + x^2}}}{1} = \lim_{x \to +\infty} \frac{x}{\sqrt{1 + x^2}}$$

$$= \lim_{x \to +\infty} \frac{1}{\frac{x}{\sqrt{1 + x^2}}} = \lim_{x \to +\infty} \frac{\sqrt{1 + x^2}}{x}, \quad \text{etc.,}$$

while elementary algebraic manipulations easily yield

$$\lim_{x \to +\infty} \frac{\sqrt{1 + x^2}}{x} = \lim_{x \to +\infty} \sqrt{\frac{1}{x^2} + 1} = 1.$$

8.7 Tests for Increase and Decrease of a Function on a Closed Interval and at a Point

Definitions. We say that a function $f(x)$ defined on a closed interval $[a, b]$ is *nondecreasing* on $[a, b]$ if given any x_1 and x_2 in $[a, b]$, the condition $x_1 < x_2$ implies that $f(x_1) \leqslant f(x_2)$.

If $x_1 < x_2$ always implies that $f(x_1) < f(x_2)$ then $f(x)$ is said to be *increasing* on $[a, b]$.

We say that a function $f(x)$ is *nonincreasing* on $[a, b]$ if given any x_1 and x_2 in $[a, b]$, $x_1 < x_2$ implies that $f(x_1) \geqslant f(x_2)$ and $f(x)$ is *decreasing* on $[a, b]$ if $x_1 < x_2$ implies that $f(x_1) > f(x_2)$.

A function $f(x)$ is said to be *monotone* on $[a, b]$ if $f(x)$ is only *nondecreasing*, in particular increasing, on $[a, b]$ or only *nonincreasing*, in particular decreasing, on $[a, b]$.

Theorem 8.12. *Let $f(x)$ be continuous on $[a, b]$ and let $f(x)$ have a derivative $f'(x)$ on (a, b). For the function $f(x)$ to be nondecreasing on $[a, b]$ it is necessary and sufficient that $f'(x) \geqslant 0$ for all x in (a, b).*

Necessity. Let $f(x)$ be nondecreasing on $[a, b]$ (Fig. 8.21). We prove that $f'(x) \geqslant 0$ on (a, b).

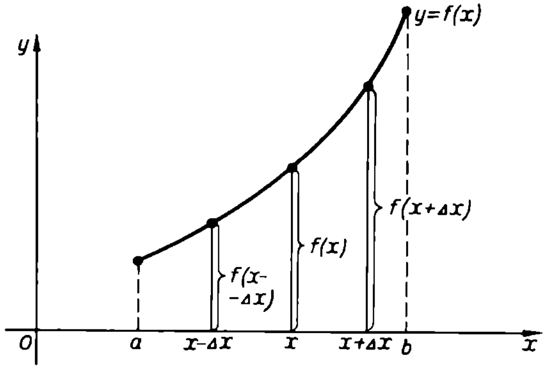


Fig. 8.21

◄ Consider two points x and $x + \Delta x$ in (a, b). Since $f(x)$ is nondecreasing then Δx and $f(x + \Delta x) - f(x)$ are of the same sign for any Δx so that

$$\frac{f(x + \Delta x) - f(x)}{\Delta x} \geqslant 0.$$

Notice that at any point x in (a, b) there exists a derivative $f'(x)$. Then from the above inequality it follows that

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \geqslant 0.$$

Hence at any x in (a, b) $f'(x) \geqslant 0$. ►

Sufficiency. Let $f'(x) \geqslant 0$ on (a, b). We prove that $f(x)$ is nondecreasing on $[a, b]$.

◄ Indeed, let $x_1 < x_2$ be any two points in $[a, b]$. By Theorem 8.7 we have

$$f(x_2) - f(x_1) = f'(\xi)(x_2 - x_1),$$

where $x_1 < \xi < x_2$.

Since $f'(x) \geqslant 0$ at any point x in (a, b) then $f'(\xi) \geqslant 0$. Also, $x_2 > x_1$. This implies that

$$f(x_2) \geqslant f(x_1).$$

Hence, $f(x)$ is nondecreasing on $[a, b]$ since $f(x_1) \leqslant f(x_2)$ for $x_1 < x_2$. ►

Similarly, we can prove the following.

Theorem 8.13. *Let $f(x)$ be continuous on $[a, b]$ and let $f(x)$ have a derivative $f'(x)$ on (a, b). For the function $f(x)$ to be nonincreasing on $[a, b]$ it is necessary and sufficient that $f'(x) \leqslant 0$ for all x in (a, b).*

Therefore if $f'(x)$ does not change its sign on some interval, $f(x)$ is monotone on this interval. Also true is the following proposition: if $f'(x) > 0$ in (a, b) then $f(x)$ increases on $[a, b]$. This proposition gives the sufficient condition for a function to increase.

It is worth mentioning that if $f(x)$ increases on $[a, b]$ then it does not follow that $f'(x) > 0$ everywhere in (a, b).

Example. The function $f(x) = x^3$ increases on $[-1, 1]$. However the derivative $f'(x) = 3x^2$ is equal to zero at $x = 0$.

We can also speak of a function decreasing or increasing at a point.

A function $f(x)$ is said to be *increasing at a point* $x = x_0$ if there exists a neighbourhood $(x_0 - \delta, x_0 + \delta)$ of x_0 such that $f(x) < f(x_0)$ whenever $x < x_0$ and $f(x) > f(x_0)$ whenever $x > x_0$ (Fig. 8.22).

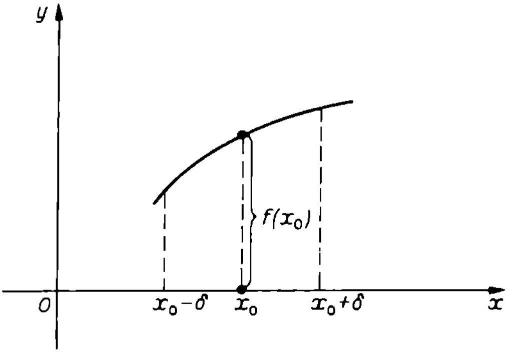


Fig. 8.22

Analogously, a function $f(x)$ is said to be *decreasing at a point* $x = x_0$ if given some neighbourhood of x_0, $f(x) > f(x_0)$ whenever $x < x_0$ and $f(x) < f(x_0)$ whenever $x > x_0$.

The following theorem specifies the *sufficient conditions for a function to be increasing or decreasing at a point.*

Theorem 8.14. *Let $f(x)$ have a derivative $f'(x_0)$ at x_0. If $f'(x_0) > 0$ then $f(x)$ increases at x_0 and if $f'(x_0) < 0$ then $f(x)$ decreases at x_0.*

◄ Let $f'(x_0) > 0$. Then

$$\lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} > 0.$$

This means that there exists $\delta > 0$ such that for all Δx

$$\frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} > 0$$

whenever $0 < |\Delta x| < \delta$.

Whence it follows that if $0 < |\Delta x| < \delta$, Δx and $f(x_0 + \Delta x) - f(x_0)$ are of the same sign, namely, if $\Delta x < 0$ then $f(x_0 + \Delta x) - f(x_0) < 0$, i.e., $f(x_0 + \Delta x) < f(x_0)$, and if $\Delta x > 0$ then $f(x_0 + \Delta x) - f(x_0) > 0$, i.e., $f(x_0 + \Delta x) > f(x_0)$.

By definition this means that $f(x)$ increases at x_0.

Using similar reasoning, we can show that if $f'(x_0) < 0$ then $f(x)$ decreases at x_0. ►

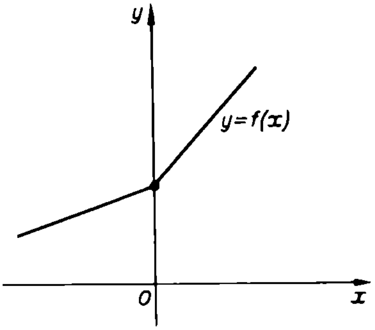


Fig. 8.23

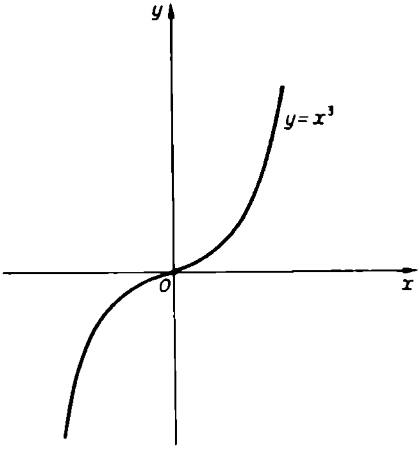


Fig. 8.24

Notice that the function shown in Fig. 8.23 increases at $x = 0$; however the derivative of this function does not exist at $x = 0$.

The function $f(x) = x^3$ increases at $x = 0$ and its derivative $f'(x) = 3x^2$ vanishes at $x = 0$ (Fig. 8.24).

8.8 Extrema of a Function. Maximum and Minimum of a Function on a Closed Interval

Local extrema. Let $f(x)$ be defined in a neighbourhood of a point x_0 and at x_0. We say that $f(x)$ has a *local maximum* at x_0 if there exists $\delta > 0$ such that

$$\Delta f = f(x) - f(x_0) \leqslant 0$$

for all x in $(x_0 - \delta, x_0 + \delta)$ (Fig. 8.25).

A function $f(x)$ is said to have a *local minimum* at x_0 if there exists $\delta > 0$ such that

$$\Delta f = f(x) - f(x_0) \geqslant 0$$

for all x in $(x_0 - \delta, x_0 + \delta)$ (Fig. 8.26).

The point x_0 at which $f(x)$ has a local maximum (minimum) is called the *point of local maximum* (*minimum*). The local maximum and local minimum of a function are called the *local extrema* of this function.

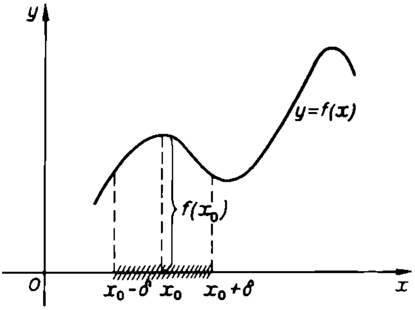


Fig. 8.25

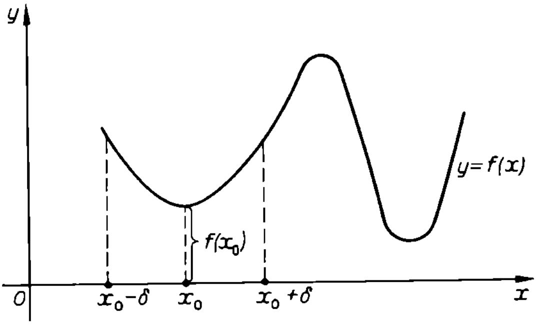


Fig. 8.26

These definitions mean that $f(x_0)$ is a local maximum of $f(x)$ if there exists an open interval $(x_0 - \delta, x_0 + \delta)$ such that $f(x_0)$ is a maximum of $f(x)$ on this interval and $f(x_0)$ is a local minimum of $f(x)$ if there exists an open interval $(x_0 - \delta, x_0 + \delta)$ such that $f(x_0)$ is a minimum of $f(x)$ on this interval.

The term "local extremum" is used here because we consider a maximum or minimum of a function within some neighbourhood of a given point rather than over the domain of a function. For example, the function $y = f(x)$ shown in Fig. 8.27 has a local maximum at x_0 and a local minimum at x_1 but $f(x_0) < f(x_1)$.

In what follows we shall speak of extrema of a function omitting "local" for brevity.

A function $f(x)$ is said to have a *strict maximum* (*minimum*) at a point x_0 if there exists $\delta > 0$ such that

$$f(x) - f(x_0) < 0 \quad (f(x) - f(x_0) > 0)$$

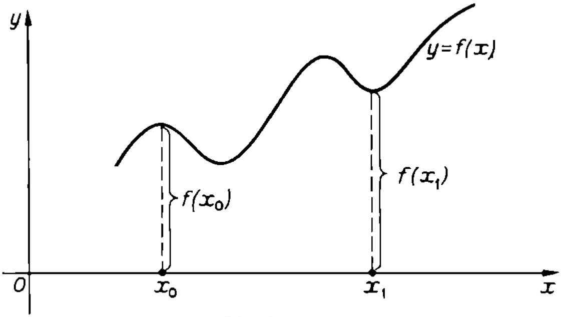


Fig. 8.27

whenever $0 < |x - x_0| < \delta$. Then the point x_0 is called the *point of strict maximum* (*minimum*) of a function. Here we do not assume that $f(x)$ is continuous at x_0. For example, the function

$$f(x) = \begin{cases} x^2 & \text{for } x \neq 0 \\ 1 & \text{for } x = 0 \end{cases}$$

is not continuous at $x = 0$ but $f(x)$ has a maximum at $x = 0$. Indeed, there exists $\delta > 0$, say $\delta = 1$, such that $f(x) - f(0) = f(x) - 1 < 0$ for all $x \neq 0$ in $(-1, 1)$ (Fig. 8.28).

Problems. **(1)** Using the definitions of maximum and minimum, prove that the function

$$f(x) = \begin{cases} e^{-1/x^2} & \text{for } x \neq 0 \\ 0 & \text{for } x = 0 \end{cases}$$

has a minimum at $x = 0$, and the function

$$g(x) = \begin{cases} xe^{-1/x^2} & \text{for } x \neq 0 \\ 0 & \text{for } x = 0 \end{cases}$$

has no extremum at $x = 0$.

(2) Find out whether the function $f(x) = (x - x_0)^n \varphi(x)$ has a maximum or minimum or has no extremum at x_0 provided that (1) $\varphi(x)$ is continuous at x_0 and $\varphi(x_0) \neq 0$, (2) the derivative $\varphi'(x)$ does not exist at x_0 and (3) n is natural.

Theorem 8.15 (necessary condition for extremum). *A function $f(x)$ can have an extremum only at the points where its derivative $f'(x)$ either is equal to zero or does not exist.*

◄ Let $f(x)$ have a derivative at x_0 and let $f'(x_0) \neq 0$. For definiteness we set $f'(x_0) > 0$. Then $f(x)$ is an increasing function at x_0 so that there exists $\delta > 0$ such that $f(x) < f(x_0)$ for all x in $(x_0 - \delta, x_0)$ and $f(x_0) < f(x)$ for all x in $(x_0, x_0 + \delta)$ (Fig. 8.29). This implies that there exists no neighbourhood of x_0 where $f(x_0)$ is a maximum or a minimum of $f(x)$ and the point x_0 is neither a point of maximum nor a point of minimum of $f(x)$.

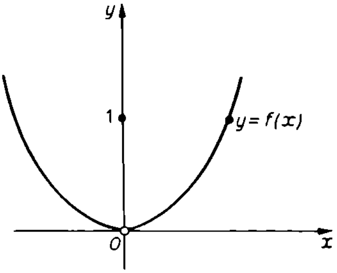


Fig. 8.28

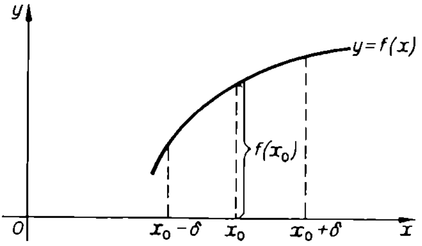


Fig. 8.29

Using similar reasoning we arrive at the same conclusion for $f'(x_0) < 0$.

Hence, if there exists a derivative $f'(x)$ at a point x_0 and if $f'(x_0) \neq 0$ then $f(x)$ can have neither a maximum nor a minimum at x_0. Thus $f(x)$ can have an extremum only at a point where its derivative $f'(x)$ either is equal to zero or does not exist. ►

Figure 8.30 offers a geometric illustration of Theorem 8.15. The function $y = f(x)$ has extrema at the points x_1, x_2, x_3 and x_4. The derivative $f'(x)$ does not exist at x_1 and x_4 and $f'(x)$ is equal to zero at x_2 and x_3.

The points where the necessary condition for extremum of a function $f(x)$ is satisfied are sometimes called the *critical points* of $f(x)$. These are the roots of the equation $f'(x) = 0$ and the points where $f'(x)$ does not exist (in particular, the points where $f'(x)$ is an infinity). The points where $f'(x) = 0$ are called the *stationary points* of $f(x)$, since the rate of change of the function $f(x)$ at these points is zero.

Theorem 8.15 specifies only the necessary condition for extremum of a function $f(x)$ and it does not necessarily have a maximum or a minimum at every critical point. For example, if $f(x) = x^3$ then $f'(0) = 0$ so that the point $x = 0$ is a critical point. However $f(x) = x^3$ has no extremum at $x = 0$ since $f(0) = 0$ and $f(x) < 0$ for $x < 0$ and $f(x) > 0$ for $x > 0$, which means that $f(x)$ increases at $x = 0$.

The following theorems specify the *sufficient conditions for a function to have a maximum or a minimum at a point.*

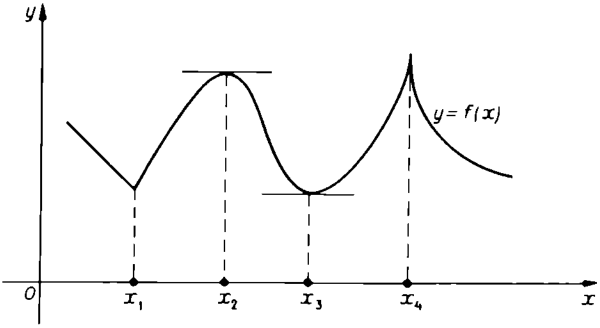


Fig. 8.30

Theorem 8.16. *Let $x = x_0$ be a critical point of a function $f(x)$ where either $f'(x_0) = 0$ or $f'(x_0)$ does not exist and let $f(x)$ be continuous at x_0. Suppose that there exists $\delta > 0$ such that $f'(x) > 0$ for all x in $(x_0 - \delta, x_0)$ and $f'(x) < 0$ for all x in $(x_0, x_0 + \delta)$ so that the derivative $f'(x)$ changes its sign from positive to negative at x_0 when x moves through x_0 from left to right. Then $f(x)$ has a maximum at x_0.*

◄ Since $f'(x) > 0$ on $(x_0 - \delta, x_0)$ then $f(x)$ increases on the closed interval $[x_0 - \delta, x_0]$; similarly, $f'(x) < 0$ on $(x_0, x_0 + \delta)$ implies that $f(x)$ decreases on the closed interval $[x_0, x_0 + \delta]$. Hence $f(x_0)$ is a maximum of $f(x)$ within the neighbourhood $(x_0 - \delta, x_0 + \delta)$ of x_0 (Fig. 8.31). This means that $f(x_0)$ is a local maximum of $f(x)$. ►

Theorem 8.17. *Let $x = x_0$ be a critical point of a function $f(x)$ where either $f'(x_0) = 0$ or $f'(x_0)$ does not exist and let $f(x)$ be continuous at x_0. Suppose that there exists $\delta > 0$ such that $f'(x) < 0$ for all x in $(x_0 - \delta, x_0)$ and $f'(x) > 0$ for all x in $(x_0, x_0 + \delta)$ so that the derivative $f'(x)$ changes its sign from negative to positive at x_0 when x moves through x_0 from left to right. Then $f(x)$ has a minimum at x_0.*

The proof of Theorem 8.17 is similar to that of Theorem 8.16.

If within some neighbourhood $(x_0 - \delta, x_0 + \delta)$ of a critical point x_0 the derivative $f'(x)$ is of the same sign on the left and on the right of x_0 the function $f(x)$ has no extremum at x_0. Thus, if $f'(x) > 0$ for all x in

$(x_0 - \delta, x_0)$ and for all x in $(x_0, x_0 + \delta)$, then given any whatever small $(x_0 - \delta, x_0 + \delta)$ $f(x)$ increases both on the left and on the right of x_0 so that $f(x_0)$ is neither a maximum nor a minimum of $f(x)$ in $(x_0 - \delta, x_0 + \delta)$, i.e., $f(x)$ has neither a maximum nor a minimum at x_0.

For the sufficient conditions given by Theorem 8.16 and Theorem 8.17 to be satisfied it is important that the function $f(x)$ be continuous at x_0. For example, if

$$f(x) = \begin{cases} -x & \text{for } x < 0, \\ x + 1 & \text{for } x \geqslant 0, \end{cases}$$

then the derivative $f'(x)$ does not exist at $x = 0$ (Fig. 8.32). The derivative $f'(x)$ changes its sign when x moves through $x = 0$; however, $f(x)$ has no extremum at $x = 0$ since there exists no neighbourhood of $x = 0$ where $f(0) = 1$ would be either a maximum or a minimum of $f(x)$. The point is that the function $f(x)$ is not continuous at $x = 0$.

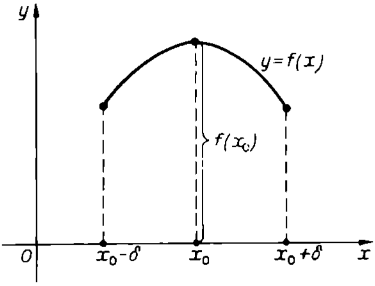


Fig. 8.31

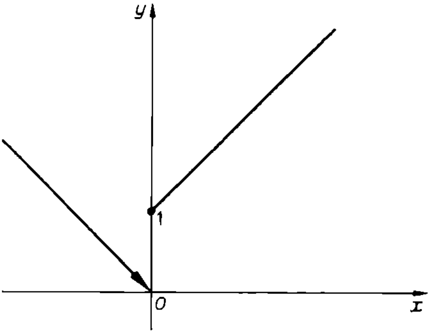


Fig. 8.32

The general procedure of computing extrema of a function involves the following steps:

(a) Compute the derivative $f'(x)$ and find the roots of the equation $f'(x) = 0$;

(b) Find all points where $f'(x)$ does not exist. These points along with the roots of $f'(x) = 0$ are critical points of $f(x)$.

(c) Determine the signs of $f'(x)$ on the left and on the right of every critical point. Then $f(x)$ has a maximum at a critical point x_0 if $f'(x)$ changes its sign from positive to negative when x moves through this critical point from left to right and $f(x)$ has a minimum at x_0 if $f'(x)$ changes its sign from negative to positive at x_0 when x moves through this point from left to right. If $f'(x)$ does not change its sign when x moves through x_0 $f(x)$ has neither a maximum nor a minimum at x_0.

Examples. (1) Investigate the function $y = x^2 e^{-x}$ for extremum.

◄ (a) Computing the derivative y' of y, we have $y' = 2xe^{-x} - x^2 e^{-x} = e^{-x}(2 - x)x$.

(b) Solving the equation $y' = 0$ gives the critical points $x = 0$ and $x = 2$.

(c) Determining the signs of the derivative, we conclude that: (i) y' is negative on the left of $x = 0$ and positive on the right of $x = 0$; (ii) y' is positive on the left of $x = 2$ and negative on the right of $x = 2$. Hence, $f(x)$ has a maximum at $x = 2$ and a minimum at $x = 0$ (Fig. 8.33). ►

(2) Investigate the function $y = x^{2/3}$ for extremum.

◄ (a) The derivative $y' = \frac{2}{3}\frac{1}{\sqrt[3]{x}}$.

(b) The derivative y' does not vanish; however it does not exist at $x = 0$ so that $y'(x) \to \infty$ as $x \to 0$. Hence the only critical point is $x = 0$.

(c) The derivative $y'(x)$ is negative on the left of $x = 0$ and positive on the right of $x = 0$. Hence the function $f(x)$ has a minimum at $x = 0$ (Fig. 8.34). ►

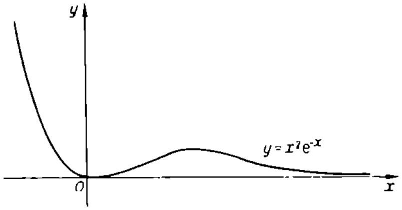


Fig. 8.33

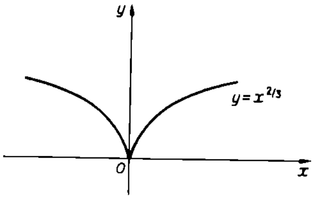


Fig. 8.34

(3) Investigate the function $y = x^3$ for extremum.

◄ (a) The derivative $y' = 3x^2$.

(b) The critical point as a solution of $3x^2 = 0$ is $x = 0$.

(c) The derivative $y'(x)$ is positive on the left and on the right of $x = 0$. Hence the function $f(x)$ increases at $x = 0$ and has neither a maximum nor a minimum at $x = 0$. ►

Remark. If a function $f(x)$ has a minimum at a point x_0 this does not imply that $f(x)$ necessarily increases on the right of x_0 and decreases on the left of x_0. Consider the function

$$f(x) = \begin{cases} x^2\left(2 - \sin\frac{1}{x}\right) & \text{for } x \neq 0, \\ 0 & \text{for } x = 0. \end{cases}$$

It is easy to see that at $x = 0$ $f(x)$ is continuous and has a minimum (Fig. 8.35). The derivative $f'(x) = 2x\left(2 - \sin\frac{1}{x}\right) + \cos\frac{1}{x}$ is continuous in any neighbourhood of $x = 0$ except at $x = 0$ and $f'(x)$ changes its sign infinitely many times. The function $f(x)$ is not monotone on the left and on the right of $x = 0$.

Applying the second derivative to investigation of a function for extremum. The following theorem specifies the sufficient conditions for a function to have an extremum at a point.

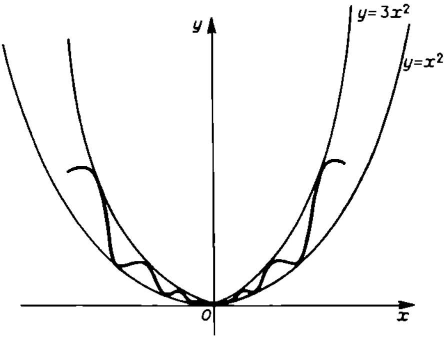


Fig. 8.35

Theorem 8.18. *Let $f(x)$ have the first and second derivatives at a point x_0 and let $f'(x_0) = 0$ and $f''(x_0) \neq 0$. Then at x_0 the function $f(x)$ has a maximum if $f''(x_0) < 0$ and a minimum if $f''(x_0) > 0$.*

◄ Observe that the point x_0 is a critical point of the function $f(x)$ since $f'(x_0) = 0$. Let $f''(x_0) < 0$. This implies that at x_0 the first derivative $f'(x)$ of $f(x)$ decreases so that there exists a neighbourhood $(x_0 - \delta, x_0 + \delta)$ of x_0 such that $f'(x) > f'(x_0) = 0$ for all x in $(x_0 - \delta, x_0)$ and $f'(x) < f'(x_0) = 0$ for all x in $(x_0, x_0 + \delta)$. Hence $f'(x)$ changes its sign from positive to negative at x_0 when x moves through x_0 from left to right. Thus $f(x)$ has a maximum at x_0.

Similar reasoning yields that $f(x)$ has a minimum at x_0 if $f''(x_0) > 0$ at x_0. ►

Theorem 8.18 enables us to lay down the following useful procedure for investigating a function for extremum. First, we find all critical points of a function as outlined in the general procedure given above. Second, we compute the second derivative $f''(x)$ at a critical point and determine its sign if $f''(x)$ exists. If at some critical point x_0 $f''(x_0) < 0, f(x)$ has a maximum at x_0 and if $f''(x_0) > 0$ at x_0 then $f(x)$ has a minimum at x_0. If the second derivative either is equal to zero or does not exist at x_0 then we can decide on the extremum of $f(x)$ at this point by using the first derivative of $f(x)$ as specified by the general procedure.

Example. Investigate the function $y = e^{-x^2}$ for extremum.

◄ We have $y' = -2xe^{-x^2}$ so that $x = 0$ is a critical point of $f(x)$. We find $y'' = -2e^{-x^2} + 4x^2e^{-x^2}$ and $y''(0) = -2 < 0$ so that $f(x)$ has a maximum at $x = 0$ (Fig. 8.36). ►

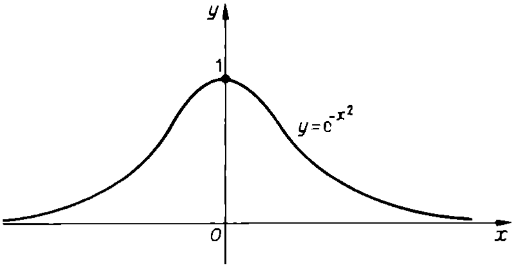


Fig. 8.36

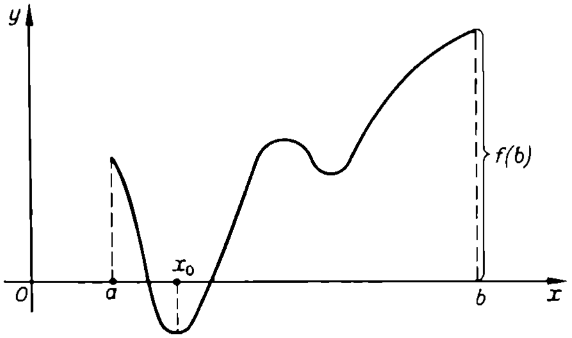


Fig. 8.37

The absolute maximum and minimum of a function continuous on a closed interval. If a function $f(x)$ is defined and continuous on a closed interval $[a, b]$ then, by virtue of Theorem 7.30, $f(x)$ attains its absolute maximum and absolute minimum on $[a, b]$.

If $f(x)$ attains its maximum M at an interior point x_0 of $[a, b]$, i.e., $a < x_0 < b$, then $M = f(x_0)$ is a local maximum of $f(x)$ since in this case there exists a neighbourhood of x_0 such that at all points belonging to this neighbourhood which are on the left and on the right of x_0 the values of $f(x)$ do not exceed $f(x_0)$. However, the function $f(x)$ can attain its absolute maximum M at the endpoints of $[a, b]$. So, if we wish to find an absolute maximum of a function $f(x)$ continuous on a closed interval $[a, b]$ we have to find all maxima of $f(x)$ in an open interval (a, b) and the values of $f(x)$ at the endpoints of $[a, b]$, i.e., $f(a)$ and $f(b)$, and choose the greatest value as the absolute maximum of $f(x)$ on $[a, b]$.

The absolute minimum of a function $f(x)$ continuous on $[a, b]$ is the least value of $f(x)$ among all the minima of $f(x)$ in (a, b) and the values of $f(a)$ and $f(b)$. In the case of a function shown in Fig. 8.37 we have $M = f(b)$ and $m = f(x_0)$.

Example. Suppose that we have a square sheet of steel whose side is a and we wish to make a box of maximal volume by cutting out four equal squares at the vertices of the given sheet as shown in Fig. 8.38 and flanging the sheet. How should we choose the size of the cut out squares to get a box of maximal volume?

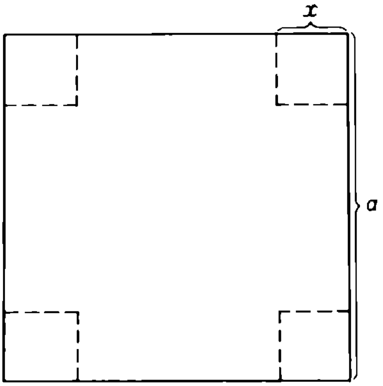


Fig. 8.38

◄ The volume of a box being a function of x is given by

$$v(x) = x(a - 2x)^2, \quad 0 \leqslant x \leqslant \frac{a}{2}.$$

The first derivative of $v(x)$ is

$$\frac{dv}{dx} = (a - 2x)^2 - 4x(a - 2x) = (a - 2x)(a - 6x)$$

so that the critical points of $v(x)$ are $x_1 = a/6$ and $x_2 = a/2$.

The open interval $(0, a/2)$ contains the critical point $x_1 = a/6$.

The second derivative of $v(x)$ is

$$\frac{d^2v}{dx^2} = -2(a - 6x) - 6(a - 2x).$$

We have $\frac{d^2v}{dx^2} < 0$ at $x = \frac{a}{6}$ so that at this point the function $v(x)$ has the maximum

$$v\left(\frac{a}{6}\right) = \frac{a}{6}\left(\frac{2}{3}a\right)^2 = \frac{2a^3}{27}.$$

At the endpoints of $[0, a/2]$ we have $v(0) = v(a/2) = 0$.

Therefore, we get the maximum of $v(x)$, i.e., the box of maximal volume, by cutting out four equal squares with side $x = a/6$ from the given sheet of steel. In this case the maximal volume of the box is $2a^3/27$. ►

8.9 Investigating the Shape of a Curve. Points of Inflection

Convexity of a curve and points of inflection. Let a curve be specified by a function $y = f(x)$ and let $y = f(x)$ have a finite derivative $f'(x_0)$ at a point x_0 so that at the point $M_0(x_0, f(x_0))$ of the curve there exists a tangent which is not parallel to the y-axis.

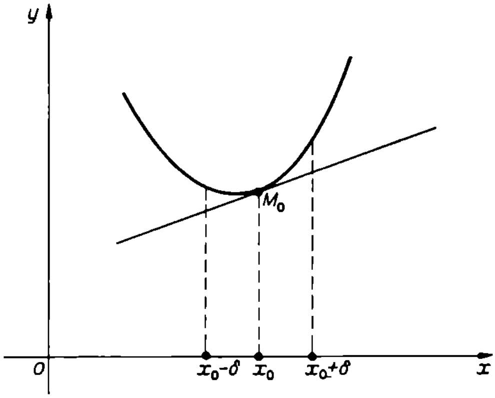


Fig. 8.39

We say that a curve is *convex downward* at a point M_0 if there exists a neighbourhood $(x_0 - \delta, x_0 + \delta)$ of a point x_0 such that all points of the curve with abscissas contained in $(x_0 - \delta, x_0 + \delta)$ lie above the tangent to this curve at M_0 (Fig. 8.39).

A curve is said to be *convex upward* at a point M_0 if all points of this curve with abscissas in some neighbourhood of x_0 lie below the tangent to the curve at M_0 (Fig. 8.40).

Let $y = f(x)$ be a function differentiable on an open interval (a, b). We say that the graph of $y = f(x)$ is *convex upward* (*downward*) on (a, b) if the graph does not lie above (below) the tangent to $y = f(x)$ for any x in (a, b).

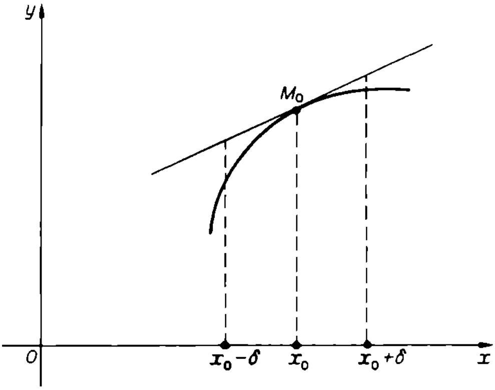


Fig. 8.40

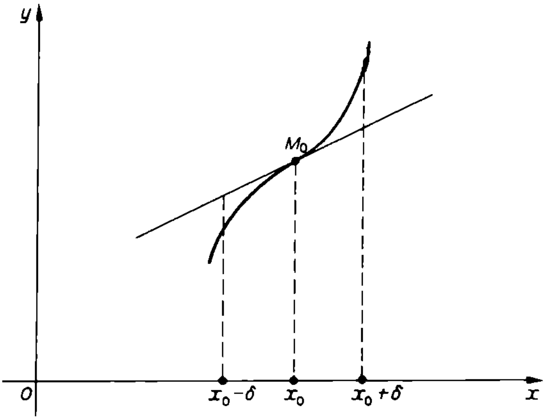


Fig. 8.41

The point $M_0(x_0, f(x_0))$ is called the *point of inflection* of a curve $y = f(x)$ if there exists a neighbourhood $(x_0 - \delta, x_0 + \delta)$ of x_0 such that for all x in $(x_0 - \delta, x_0 + \delta)$ the curve is convex upward if $x < x_0$ and convex downward if $x > x_0$ or vice versa (Fig. 8.41). In other words, the point M_0 is the point of inflection if on the left and on the right of M_0 the curve lies on different sides of the tangent at M_0.

Now we shall consider an analytical procedure for investigating the sense of convexity and points of inflection of a curve.

Let us choose a point on a curve $y = f(x)$ and a point on the tangent to $y = f(x)$ at $M_0(x_0, f(x_0))$. Suppose that these points have the same abscissa x and y is the ordinate of the point chosen on the curve while Y is the ordinate of the point on the tangent (Fig. 8.42). Evidently, if $y - Y > 0$ for all $x \neq x_0$ in a sufficiently small neighbourhood of x_0 the curve is convex downward at M_0 and if $y - Y < 0$ for all $x \neq x_0$ in a sufficiently small neighbourhood of x_0 the curve is convex upward at M_0. Hence, to determine the sense of convexity of a curve at M_0 it suffices to investigate the sign of the difference $y - Y$ in a neighbourhood of x_0.

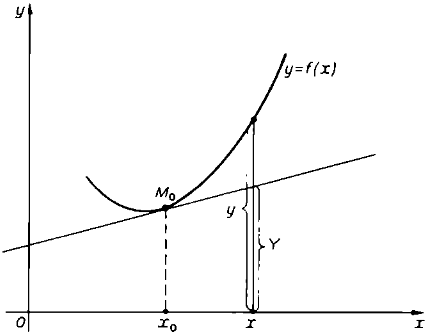


Fig. 8.42

Since the equation of the tangent to the curve is

$$Y - f(x_0) = f'(x_0)(x - x_0)$$

we have

$$\begin{aligned} y - Y &= f(x) - [f(x_0) + f'(x_0)(x - x_0)] \\ &= [f(x) - f(x_0)] - f'(x_0)(x - x_0) \end{aligned}$$

or

$$\Delta = y - Y = [f(x_0 + h) - f(x_0)] - f'(x_0)h,$$

where $h = x - x_0$.

Let $f(x)$ have the second derivative $f''(x)$ at x_0 and in some neighbourhood of x_0. Applying the mean value theorem (Theorem 8.7) to the above identity, we get

$$\Delta = f'(x_0 + \theta h)h - f'(x_0)h = [f'(x_0 + \theta h) - f'(x_0)]h,$$

where $\theta = \theta(h)$ and $0 < \theta < 1$.

Since $f''(x)$ exists at x_0 then

$$\lim_{\Delta x \to 0} \frac{f'(x_0 + \Delta x) - f'(x_0)}{\Delta x} = f''(x_0)$$

so that

$$\lim_{\theta h \to 0} \frac{f'(x_0 + \theta h) - f'(x_0)}{\theta h} = f''(x_0)$$

and

$$f'(x_0 + \theta h) - f'(x_0) = f''(x_0) \cdot \theta h + \alpha(\theta h) \cdot \theta h,$$

where $\alpha(\theta h) \to 0$ as $h \to 0$.

Then we can write

$$\Delta = y - Y = [f''(x_0) + \alpha(\theta h)]\theta h^2.$$

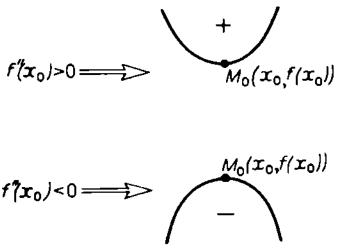


Fig. 8.43

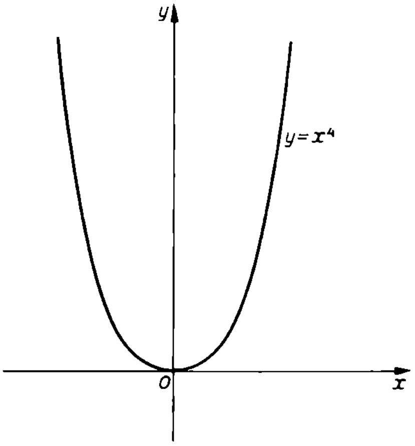


Fig. 8.44

Let $f''(x_0) \neq 0$. Since $\alpha(\theta h)$ is an infinitesimal as $h \to 0$ there exists $\delta > 0$ such that $f''(x_0) + \alpha(\theta h)$ is of the same sign as $f''(x_0)$ and $\theta h^2 > 0$ whenever $0 < |h| < \delta$. Hence, if $f''(x_0) > 0$ then $y - Y > 0$ for all x sufficiently near to x_0 and the curve $y = f(x)$ is convex downward at the point $M_0(x_0, f(x_0))$ and if $f''(x_0) < 0$ then $y = f(x)$ is convex upward at M_0 (Fig. 8.43). Whence follows the ***necessary condition*** for a point to be a point of inflection: the point $M_0(x_0, f(x_0))$ may be the point of inflection of the curve $y = f(x)$ only if $f''(x_0) = 0$ or if $f''(x_0)$ does not exist.

This condition is not a sufficient one. For example, if $f(x) = x^4$ then $f''(x) = 12x^2$ and $f''(0) = 0$; however, the point $O(0, 0)$ is not a point of inflection of $y = x^4$: at this point the curve is convex downward (Fig. 8.44).

The following theorem gives the *sufficient condition* for a point to be a point of inflection of a curve.

Theorem 8.19. *Let $f(x)$ have the second derivative in a neighbourhood of x_0 and let the second derivative of $f(x)$ be continuous at x_0. Then the point $M_0(x_0, f(x_0))$ is a point of inflection of the curve $y = f(x)$ if $f''(x_0) = 0$ and $f''(x)$ changes its sign at x_0 when x passes through x_0.*

◀ Indeed, let $f''(x_0) = 0$ and let there exist a neighbourhood $(x_0 - \delta, x_0 + \delta)$ of x_0 such that the sign of $f''(x)$ for all $x < x_0$ differs from that for all $x > x_0$. This means that the sense of convexity changes at $M_0(x_0, f(x_0))$ so that M_0 is a point of inflection.

Suppose now that $f''(x_0) = 0$ and there exists some neighbourhood of x_0 such that $f''(x)$ is of the same sign for all $x < x_0$ and all $x > x_0$. Then M_0 is not a point of inflection since the curve is convex downward at M_0 if $f''(x) > 0$ both on the left and on the right of x_0 and the curve is convex upward at M_0 if $f''(x) < 0$ both on the left and on the right of x_0. ▶

Problem. Let $f(x)$ have a finite third derivative at x_0 and let $f''(x_0) = 0$ and $f'''(x_0) \neq 0$. Prove that $M_0(x_0, f(x_0))$ is a point of inflection of the graph of $f(x)$.

Sometimes it happens that at the point of inflection $M_0(x_0, f(x_0))$ the tangent to the curve $y = f(x)$ is vertical, i.e., parallel to the y-axis, so that $f''(x)$ does not exist at x_0. For example, if $f(x) = x^{1/3}$ then $f'(x) = \frac{1}{3} x^{-2/3}$ and $f''(x) = -\frac{2}{9}\frac{1}{\sqrt[3]{x^5}}$. Clearly, there exist no points where $f''(x) = 0$ and at $x = 0$ $f''(x)$ does not exist. Let us investigate the sign of $f''(x)$ in a neighbourhood of $x = 0$. We have $f''(x) > 0$ for all x in $(-\delta, 0)$, $\delta > 0$, and $f''(x) < 0$ for all x in $(0, \delta)$ so that the curve is convex downward on the left of the point $O(0, 0)$ and convex upward on the right of $O(0, 0)$. Hence, the point $O(0, 0)$ is the point of inflection of $y = x^{1/3}$. The tangent to this curve at $O(0, 0)$ is perpendicular to the x-axis, for $\lim\limits_{x \to 0} f'(x) = +\infty$.

Finally, we can state the following sufficient condition for a point to be a point of inflection. Let $y = f(x)$ have a tangent at $M_0(x_0, f(x_0))$ which can be parallel to the y-axis. Let $f(x)$ have the second derivative continuous in a neighbourhood of x_0 except probably at x_0. If $f''(x)$ is equal to zero or does not exist at x_0 and if $f''(x)$ changes its sign when x passes through x_0 then $M_0(x_0, f(x_0))$ is the point of inflection of $y = f(x)$.

8.10 Asymptotes of a Curve

Consider a curve having an infinite branch (Fig. 8.45). An *asymptote* of an infinite branch of a curve is a line whose distance δ to a point M on the curve tends to zero as M recedes infinitely distant from the origin of coordinates.

Vertical asymptotes. A line passing parallel to the y-axis through the point $x = x_0$ on the x-axis is a *vertical* asymptote of the graph of $y = f(x)$ if at least one of the identities

$$\lim_{x \to x_0 - 0} f(x) = \pm\infty \quad \text{or} \quad \lim_{x \to x_0 + 0} f(x) = \pm\infty$$

holds.

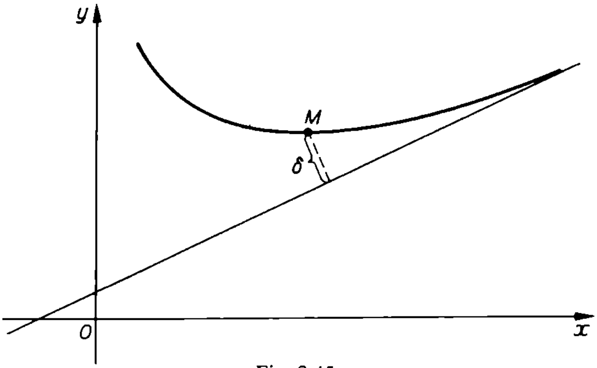


Fig. 8.45

Clearly, in this case the distance $\delta = |x - x_0|$ between the point $M(x, f(x))$ on the graph of $y = f(x)$ and the line $x = x_0$ tends to zero and M recedes infinitely distant from the origin of coordinates.

For example, the graph of $y = 1/x$ has the vertical asymptote $x = 0$ since $\lim\limits_{x \to 0-0} \dfrac{1}{x} = -\infty$ and $\lim\limits_{x \to 0+0} \dfrac{1}{x} = +\infty$ (Fig. 8.46). Similarly, the graph of $y = e^{1/x}$ has the vertical asymptote $x = 0$ for $\lim\limits_{x \to 0+0} e^{1/x} = +\infty$.

Figure 8.47 illustrates mutual positions of a curve and its vertical asymptotes.

To find vertical asymptotes of a curve $y = f(x)$ one should proceed as follows:

(1) Find discontinuities of $f(x)$ on the x-axis.

(2) Isolate all those discontinuities where at least one of the one-sided limits of $f(x)$ is equal to $+\infty$ or $-\infty$. Let these be $x_1, x_2, \ldots, x_m$. Then the lines $x = x_1, x = x_2, \ldots, x = x_m$ will be the vertical asymptotes of the

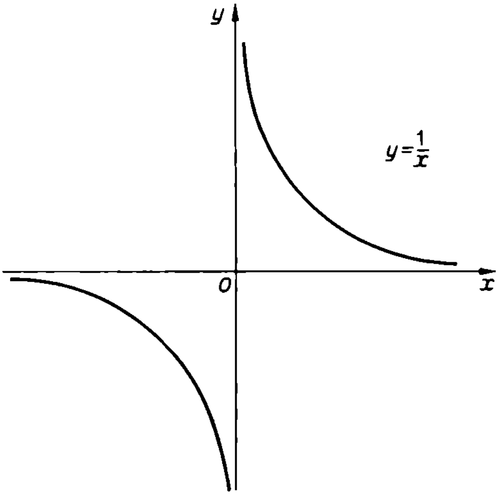


Fig. 8.46

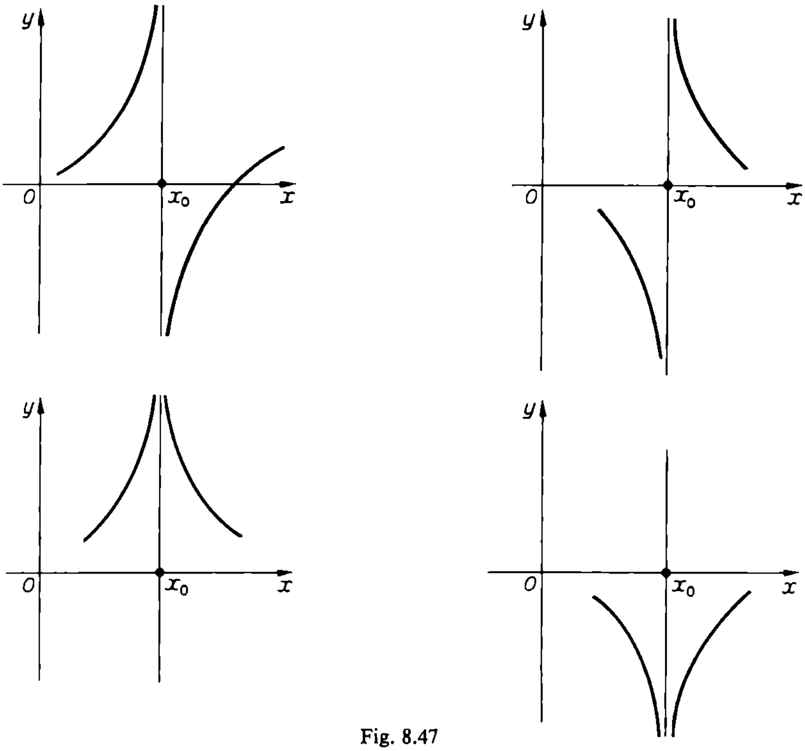


Fig. 8.47

graph of $y = f(x)$. For example, the vertical asymptotes of the graph of the curve $y = \frac{1}{x^2 - 1}$ are the lines $x = -1$ and $x = 1$ (Fig. 8.48).

The vertical line $x = x_0$ (x_0 being the endpoint of the interval where $f(x)$ is defined) is the vertical asymptote of $y = f(x)$ provided that (i) x_0 is the left end of the interval and

$$\lim_{x \to x_0 + 0} f(x) = +\infty \quad \text{or} \quad \lim_{x \to x_0 + 0} f(x) = -\infty$$

or (ii) x_0 is the right end of the interval and

$$\lim_{x \to x_0 - 0} f(x) = +\infty \quad \text{or} \quad \lim_{x \to x_0 - 0} f(x) = -\infty.$$

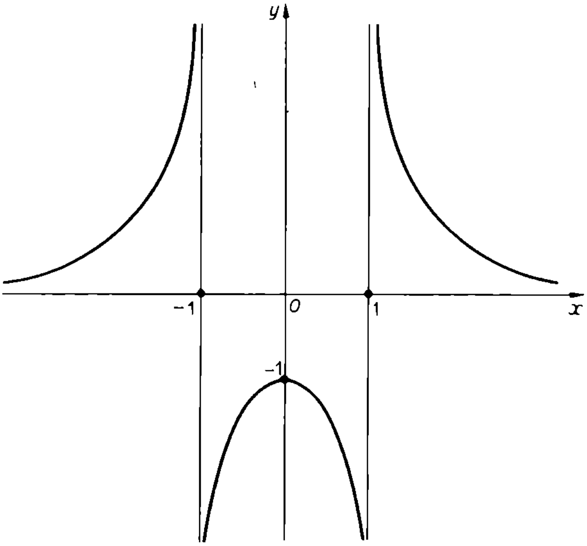


Fig. 8.48

For example, the function $y = \ln x$ is defined on the interval $0 < x < +\infty$ and $\lim_{x \to 0 + 0} \ln x = -\infty$ so that the line $x = 0$, i.e., the y-axis, is the vertical asymptote of the graph of $y = \ln x$.

Oblique asymptotes. Let $y = f(x)$ be defined for all $x \geqslant a$ ($x \leqslant a$). Suppose that the line $y = kx + b$ is an asymptote of the graph of $y = f(x)$. Then the asymptote $y = kx + b$ is called the *oblique asymptote.*

For definiteness we consider whatever large positive x. If the line $y = kx + b$ is an asymptote of $y = f(x)$ then, by definition, the distance δ from the point $M(x, f(x))$ on the curve $y = f(x)$ to the line $y = kx + b$ tends to zero as $x \to +\infty$. Let α be an angle ($\alpha \neq \pi/2$) made by the asymptote with the x-axis (Fig. 8.49). Evidently, $\delta = |MN| \cos \alpha$. Since $\cos \alpha \neq 0$ then $|MN|$ tends to zero as δ tends to zero as $x \to +\infty$ and vice versa. Noting that $|MN| = |f(x) - kx - b|$ we infer that the line $y = kx + b$ is an oblique asymptote of the graph of $y = f(x)$ if and only if

$$\lim_{x \to +\infty} (f(x) - kx - b) = 0$$

so that $f(x)$ admits a representation of the form

$$f(x) = kx + b + \alpha(x),$$

where $\lim\limits_{x \to +\infty} \alpha(x) = 0.$

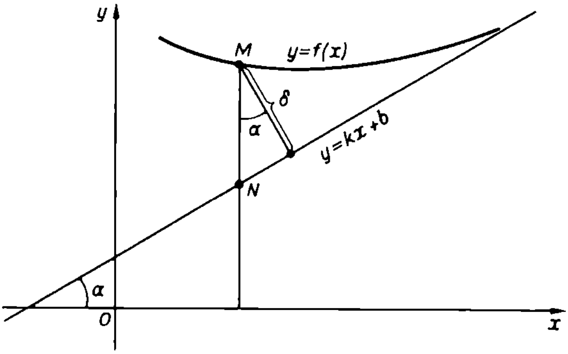


Fig. 8.49

Observe that if there exists an asymptote $y = kx + b$ of the curve $y = f(x)$ as $x \to +\infty$ then the function $y = f(x)$ is "nearly" a linear function as $x \to +\infty$, i.e., $y = f(x)$ differs from $y = kx + b$ by an infinitesimal as $x \to +\infty$.

Theorem 8.20. *For the graph of the function $y = f(x)$ to have the oblique asymptote $y = kx + b$ as $x \to +\infty$ it is necessary and sufficient that there exist both*

$$\lim_{x \to +\infty} \frac{f(x)}{x} = k \quad \text{and} \quad \lim_{x \to +\infty} [f(x) - kx] = b.$$

◄ *Necessity.* Let the graph of $y = f(x)$ have an asymptote $y = kx + b$ as $x \to +\infty$ so that

$$f(x) = kx + b + \alpha(x),$$

where $\alpha(x) \to 0$ as $x \to +\infty$.

Then

$$\lim_{x\to+\infty}\frac{f(x)}{x}=\lim_{x\to+\infty}\left[k+\frac{b}{x}+\frac{\alpha(x)}{x}\right]=k$$

and

$$\lim_{x\to+\infty}[f(x)-kx]=\lim_{x\to+\infty}[b+\alpha(x)]=b$$

i.e., both limits we are seeking for do exist.

Sufficiency. Let both limits mentioned in Theorem 8.20 exist. Since $\lim\limits_{x\to+\infty}[f(x)-kx]=b$ the difference $f(x)-kx-b$ becomes an infinitesimal as $x\to+\infty$. Denote this difference by $\alpha(x)$. Then we can write $f(x)=kx+b+\alpha(x)$ where $\alpha(x)\to0$ as $x\to+\infty$. Whence it follows that the graph of the function $y=f(x)$ has the oblique asymptote $y=kx+b$. ►

The proof is similar if $x\to-\infty$.

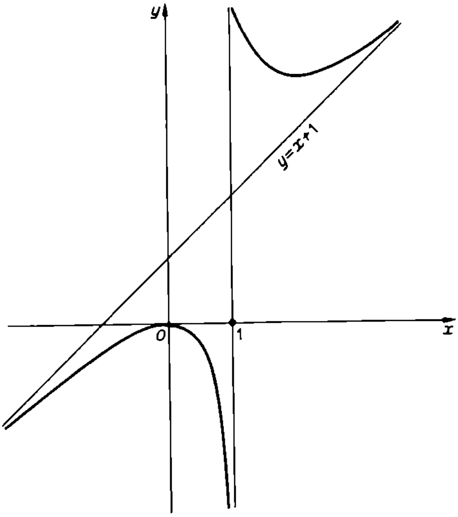


Fig. 8.50

Example. Prove that the graph of $y=\dfrac{x^2}{x-1}$ has an oblique asymptote.

◄ The graph of $y=\dfrac{x^2}{x-1}$ has the vertical asymptote $x=1$ (Fig. 8.50). Let us write this function if $f(x)=\dfrac{x^2-1+1}{x-1}=x+1+\dfrac{1}{x-1}$ where

$\frac{1}{x-1}$ tends to zero as $x \to \infty$. Hence, the function being considered can be expressed as

$$f(x) = x + 1 + \alpha(x),$$

where $\alpha(x) = \frac{1}{x-1} \to 0$ as $x \to \infty$.

Whence we conclude that the graph of the function $y = \frac{x^2}{x-1}$ has the oblique asymptote $y = x + 1$. ►

It is worth noting that the graph of $y = f(x)$ lies below its asymptote if $\Delta = f(x) - kx - b < 0$ and the graph lies above the asymptote if $\Delta > 0$.

Horizontal asymptotes. If the function $f(x)$ has the finite limit $\lim\limits_{x \to +\infty} f(x) = b$ ($\lim\limits_{x \to -\infty} f(x) = b$) as $x \to +\infty$ $(x \to -\infty)$ then the line $y = b$ is the *horizontal* asymptote of the right (left) branch of the graph of $y = f(x)$.

Clearly, a horizontal asymptote is a specific case of an oblique asymptote provided that $k = 0$.

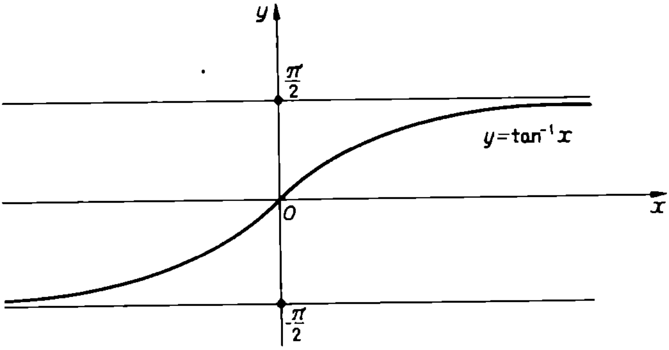


Fig. 8.51

Examples. (1) Let $y = 1/x$. Since $\lim\limits_{x \to \infty} 1/x = 0$ the graph of the function $y = 1/x$ has the horizontal asymptote $y = 0$.

(2) Let $y = \tan^{-1} x$ so that $\lim\limits_{x \to +\infty} \tan^{-1} x = \pi/2$ and $\lim\limits_{x \to -\infty} \tan^{-1} x = -\pi/2$. Hence, the right-hand branch of the graph of $y = \tan^{-1} x$ has the asymptote $y = \pi/2$ and the left-hand branch has the asymptote $y = -\pi/2$ (Fig. 8.51).

(3) Let $y = \frac{\sin x}{x}$ and $y(0) = 1$. Since $\lim_{x\to\infty} \frac{\sin x}{x} = 0$ the line $y = 0$ is the horizontal asymptote of the graph of the function $y = \frac{\sin x}{x}$ (Fig. 8.52).

As easily follows from the above example the graph of the function $y = f(x)$ can intersect its asymptote infinitely many times.

Problem. Derive the condition for asymptotes of the graph of the rational function to exist.

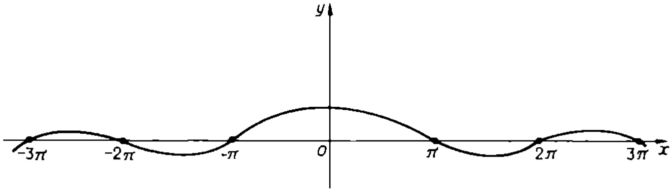


Fig. 8.52

8.11 Curve Sketching

In general the process of investigating a function and sketching its graph involves a sequence of steps which enable us to determine the basic properties of the function being considered. These are:

(a) The domain of definition of the function.

(b) The discontinuities of the function and their kinds. The vertical asymptotes.

(c) The point symmetry and the translation symmetry, i.e., whether the function is even or odd and whether the function is periodic or not.

(d) The points of intersection of the graph with the axes of coordinates.

(e) The behaviour of the function at infinity and the horizontal and oblique asymptotes.

(f) The intervals where the function is monotone and the points where the function has extrema.

(g) The intervals where the graph is convex downward and is convex upward and the points of inflection.

Based on the results of the steps (a)-(g) we can easily plot the graph of the function. To illustrate our approach to curve sketching we consider the following examples of investigating functions and constructing their graphs.

Examples. (1) $y = \frac{1}{1 + x^2}$ (Witch of Agnesi).

(a) The domain of definition is the number line.

(b) There exist no discontinuities and no vertical asymptotes.

(c) The function is even, i.e., $f(-x) = f(x)$ so that the graph is symmetric relative to the y-axis. The function is not periodic. Since the function is even it suffices to construct the graph for $x \geqslant 0$ and reflect this graph with respect to the y-axis.

(d) The graph lies above the x-axis since $y = 1$ at $x = 0$ and $y > 0$ for all x; $y \neq 0$.

(e) The graph has the horizontal asymptote $y = 0$ since $\lim\limits_{x \to \pm\infty} f(x) = \lim\limits_{x \to \pm\infty} \dfrac{1}{1 + x^2} = 0$. There are no oblique asymptotes.

(f) The first derivative of y is $y' = -\dfrac{2x}{(1 + x^2)^2}$ so that $f(x)$ increases for $x < 0$ and decreases for $x > 0$. The point $x = 0$ is critical. The derivative y' changes its sign from negative to positive when x moves through the point $x = 0$ from left to right. Hence, the function has the maximum $y(0) = 1$ at $x = 0$. This almost immediately follows from the inequality $f(x) = \dfrac{1}{1 + x^2} \leqslant 1$, holding for all x.

(g) The second derivative $y'' = -2\dfrac{1 - 3x^2}{(1 + x^2)^3}$ vanishes at the points $x = 1/\sqrt{3}$ and $x = -1/\sqrt{3}$. The graph is convex downward for $x > 1/\sqrt{3}$ and convex upward for $x < 1/\sqrt{3}$ since $y'' > 0$ in the former case and $y'' < 0$ in the latter. Hence, $x = 1/\sqrt{3} = \sqrt{3}/3$ is the point of inflection. By virtue of symmetry the point $x = -1/\sqrt{3} = -\sqrt{3}/3$ is also the point of inflection.

It is convenient to represent the results in the tabular form

Table 8.1

x	$\left(-\infty, -\frac{\sqrt{3}}{3}\right)$	$-\frac{\sqrt{3}}{3}$	$\left(-\frac{\sqrt{3}}{3}, 0\right)$	0	$\left(0, \frac{\sqrt{3}}{3}\right)$	$\frac{\sqrt{3}}{3}$	$\left(\frac{\sqrt{3}}{3}, -\infty\right)$
$f'(x)$	+	+	+	0	—	—	—
$f''(x)$	+	0	—	−2	—	0	+
$f(x)$	↗	point of inflection	↗	max	↘	point of inflection	↘

The arrows ↗ and ↘ refer to the increase and decrease of the function, respectively. The graph of the function is given in Fig. 8.53.

(2) $y = x^2 + \dfrac{1}{x}$.

(a) The domain is the number line except the point $x = 0$.

(b) The discontinuity is the point $x = 0$. We have

$$\lim_{x\to 0+0}\left(x^2+\frac{1}{x}\right) = +\infty \text{ and}$$

$$\lim_{x\to 0-0}\left(x^2+\frac{1}{x}\right) = -\infty$$

so that the line $x = 0$ is the vertical asymptote of the graph.

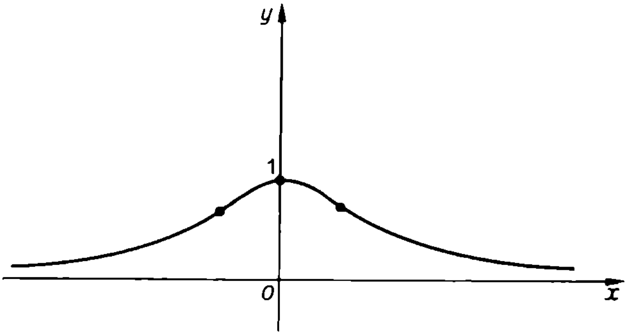


Fig. 8.53

(c) The function is neither even nor odd and it is not periodic.

(d) Putting $y = 0$, we get $x^2 + \frac{1}{x} = 0$ or $\frac{x^3+1}{x} = 0$; whence $x = -1$ so that the graph of the function meets the x-axis at the point $x = -1$.

(e) There are neither oblique nor horizontal asymptotes since

$$\lim_{x\to\pm\infty}\frac{f(x)}{x} = \lim_{x\to\pm\infty}\left(x+\frac{1}{x^2}\right) = \pm\infty.$$

(f) The first derivative is $y' = 2x - \frac{1}{x^2} = \frac{2x^3-1}{x^2}$ so that the point $x = \frac{1}{\sqrt[3]{2}}$ is the critical point. The second derivative $y'' = 2 + \frac{2}{x^3}$ is positive at the point $x = \frac{1}{\sqrt[3]{2}}$; hence, at this point the function has the minimum.

(g) The second derivative $y'' = \frac{2(x^3+1)}{x^3}$ vanishes at $x = -1$ and changes the sign from positive to negative when x passes through the point $x = -1$ from left to right. Thus the point $x = -1$ is the point of inflection. Since $y'' > 0$ for x in $(-\infty, -1)$ and for x in $(0, +\infty)$ and $y'' < 0$ on the interval $-1 < x < 0$ the graph is convex downward on the open intervals $(-\infty, -1)$ and $(0, +\infty)$ and is convex upward on the open interval $-1 < x < 0$.

The graph of the function is given in Fig. 8.54.

(3) $y = x + \dfrac{\ln x}{x}$.

(a) The domain is the positive number line $x > 0$.

(b) There are no discontinuities in the domain of definition.

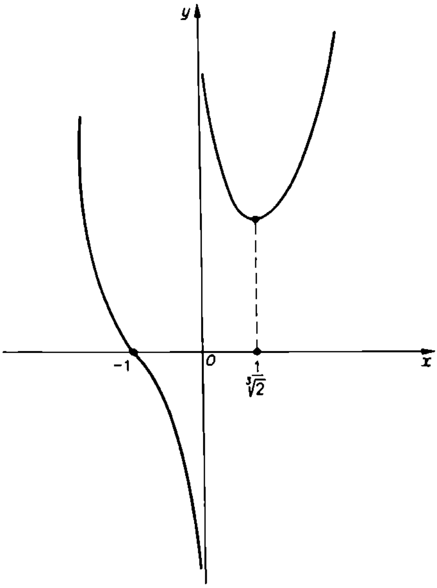


Fig. 8.54

We have $\lim\limits_{x \to 0+0} \left(x + \dfrac{\ln x}{x}\right) = -\infty$ so that the line $x = 0$ passing through the endpoint of the domain of definition is the vertical asymptote of the graph of the function.

(c) The function is neither even nor odd nor periodic.

(d) Putting $y = 0$, we have $x + \dfrac{\ln x}{x} = 0$ and $x^2 + \ln x = 0$. The solutions of these equations which can be obtained from the graph shown in Fig. 8.55 define the point of intersection of the graph under investigation with the x-axis.

(e) Since

$$\lim_{x \to +\infty} \frac{f(x)}{x} = \lim_{x \to +\infty} \left(1 + \frac{\ln x}{x^2}\right) = 1 = k$$

and

$$\lim_{x \to +\infty} [f(x) - kx] = \lim_{x \to +\infty} \frac{\ln x}{x} = 0,$$

the line $y = x$ is the oblique asymptote of the graph of the function.

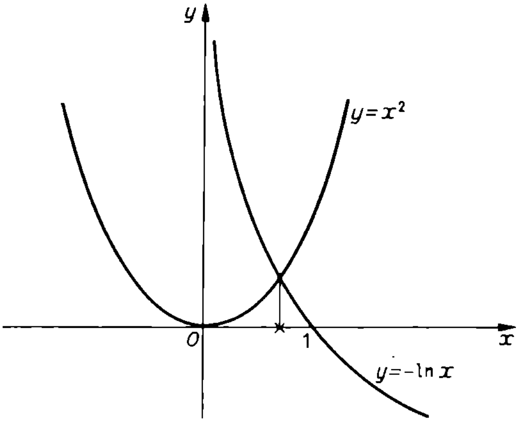


Fig. 8.55

(f) The first derivative is $y' = 1 + \frac{1}{x^2} - \frac{\ln x}{x^2} = \frac{x^2 + 1 - \ln x}{x^2}$. We can easily see from Fig. 8.56 that $x^2 + 1 > \ln x$ for all $x > 0$, Hence, $y' > 0$ for all x and $f(x)$ increases on the interval $(0, +\infty)$ and has no extrema.

(g) The second derivative $y'' = -\frac{2}{x^3} - \frac{1}{x^3} + \frac{2\ln x}{x^3} = \frac{2\ln x - 3}{x^3}$ vanishes at the point $x = e^{3/2}$ and changes the sign from negative to positive when x passes through $x = e^{3/2}$ from left to right. Hence, $x = e^{3/2}$ is the point of inflection.

The graph of the function is shown in Fig. 8.57.

(4) $y = x + \frac{1}{x^2}$.

(a) The domain is the number line except the point $x = 0$.

(b) The point $x = 0$ is the discontinuity of the second kind. The line

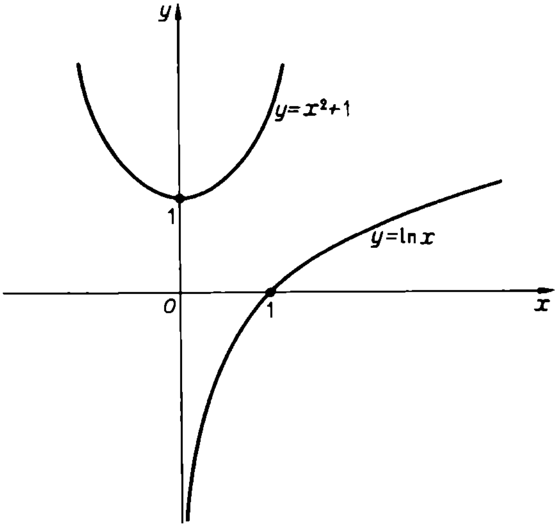


Fig. 8.56

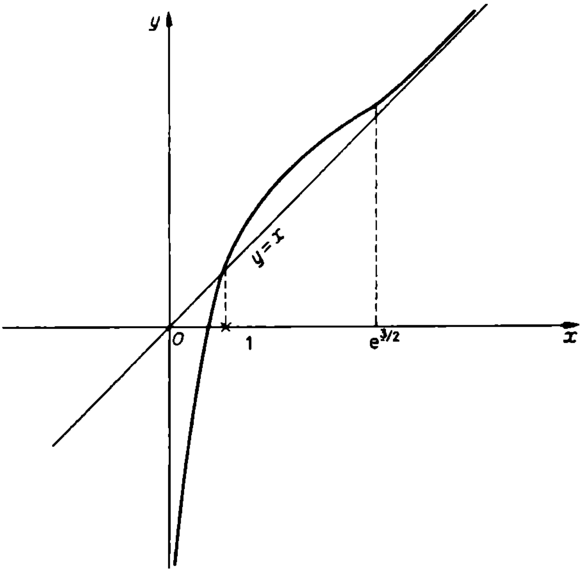


Fig. 8.57

$x = 0$ is the vertical asymptote of the graph of the function, for $\lim\limits_{x \to 0 \pm 0} \left(x + \frac{1}{x^2}\right) = +\infty$.

(c) The function is neither even nor odd nor periodic.

(d) Putting $y = 0$, we have $x^3 + 1 = 0$; whence $x = -1$ so that the graph of the function intersects the x-axis at the point $x = -1$.

(e) The graph has the oblique asymptote $y = x$ since

$$\lim_{x \to \pm\infty} \frac{f(x)}{x} = \lim_{x \to \pm\infty} \left(1 + \frac{1}{x^3}\right) = 1 = k$$

and

$$\lim_{x \to \pm\infty} [f(x) - x] = \lim_{x \to \pm\infty} \frac{1}{x^2} = 0 = b.$$

(f) The equation $y' = 1 - \frac{2}{x^3} = \frac{x^3 - 2}{x^3} = 0$ gives $x^3 - 2 = 0$ so that $x = \sqrt[3]{2}$ is the critical point of the function.

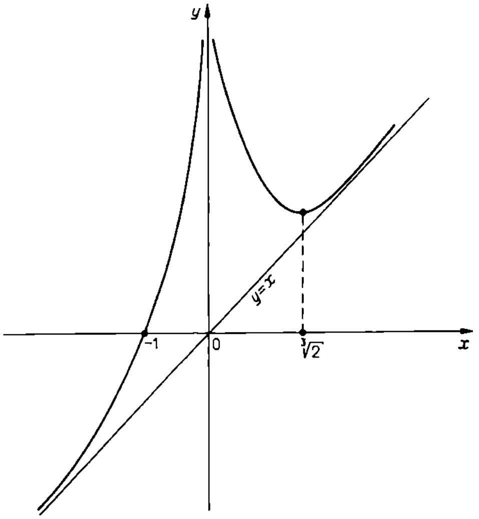


Fig. 8.58

The second derivative $y'' = 6/x^4$ is positive everywhere in the domain of definition of the function and, consequently, at $x = \sqrt[3]{2}$. Whence it follows that the function has the minimum at $x = \sqrt[3]{2}$.

(g) Since the second derivative $y'' = 6/x^4$ is positive everywhere in the domain of definition the graph is convex downward.

The graph of the function is given in Fig. 8.58.

(5) $y = \sqrt[3]{x(x-3)^2}$.

(a) The domain is the number line.

(b) The function is continuous everywhere in the domain of definition and its graph has no vertical asymptotes.

(c) The function is nonperiodic and neither even nor odd.

(d) The function vanishes at the points $x = 0$ and $x = 3$.

(e) The graph of the function has the oblique asymptote $y = x - 2$ since

$$\lim_{x \to \pm\infty} \frac{f(x)}{x} = \lim_{x \to \pm\infty} \frac{\sqrt[3]{x(x-3)^2}}{x} = 1 = k$$

and

$$\lim_{x \to \pm\infty} [f(x) - x] = \lim_{x \to \pm\infty} [\sqrt[3]{x(x-3)^2} - x]$$

$$= \lim_{x \to \pm\infty} \frac{x(x-3)^2 - x^3}{\sqrt[3]{x^2(x-3)^4} + x\sqrt[3]{x(x-3)^2} + x^2} = -2 = b.$$

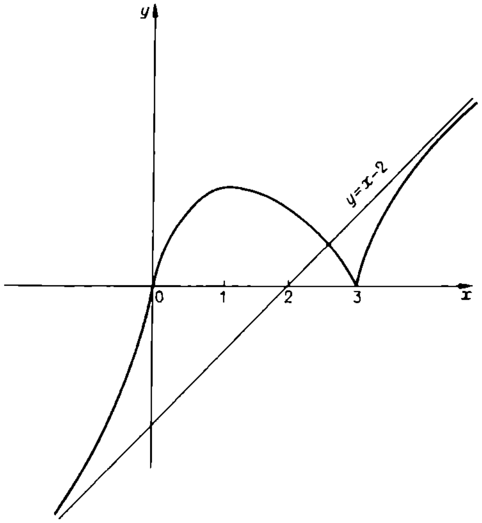


Fig. 8.59

(f) The first derivative

$$y' = \frac{(x-3)^2 + 2x(x-3)}{3[x(x-3)^2]^{2/3}} = \frac{x-1}{x^{2/3}(x-3)^{1/3}}$$

vanishes at the point $x = 1$ and does not exist at the points $x = 0$ and $x = 3$. y' is of the same sign on the left and on the right of the point $x = 0$ so that the function has no extremum at $x = 0$. On the other hand the first derivative changes its sign from positive to negative when x passes through the point $x = 1$ from left to right so that the function has a maximum at $x = 1$.

At the point $x = 3$ the function has a minimum since the first derivative changes its sign from negative to positive when x passes through the point $x = 3$ from left to right.

(g) The second derivative

$$y'' = -\frac{2}{x^{5/3}(x-3)^{4/3}}$$

does not exist at the point $x = 0$ and changes its sign from positive to negative when x passes through the point $x = 0$ from left to right so that the point $x = 0$ is the point of inflection and the graph of the function has the tangent parallel to the y-axis at (0, 0). The point $x = 3$ is not the point of inflection. The graph of the function, shown in Fig. 8.59, is convex upward in the half-plane $x > 0$.

8.12 Approximate Solution of Equations

We wish to compute the real root of the equation

$$f(x) = 0.$$

Suppose that the following conditions are satisfied:

(1) the function $f(x)$ is continuous on a closed interval $[a, b]$;

(2) the values $f(a)$ and $f(b)$ are of opposite signs so that $f(a)\cdot f(b) < 0$;

(3) there exist derivatives $f'(x)$ and $f''(x)$ on $[a, b]$ retaining their signs on $[a, b]$.

By virtue of conditions (1)-(2), Theorem 7.27 says that $f(x)$ vanishes at least at one point ξ in (a, b); this means that the equation being considered has at least one real root ξ in (a, b). Furthermore, since $f'(x)$ retains its sign on $[a, b]$ the function $f(x)$ is monotone on $[a, b]$ so that the equation has the unique real root ξ in (a, b).

We shall consider an iterative method of computing the unique real root of the equation $f(x) = 0$ with any predetermined precision.

Four cases have to be distinguished for $[a, b]$:

(a) $f'(x) > 0$ and $f''(x) > 0$;

(b) $f'(x) > 0$ and $f''(x) < 0$;

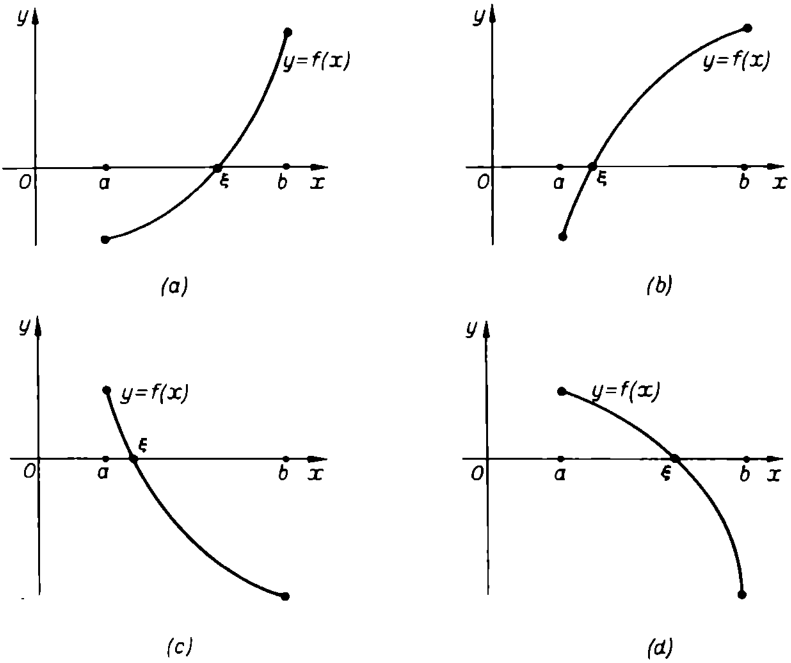
(c) $f'(x) < 0$ and $f''(x) > 0$;

(d) $f'(x) < 0$ and $f''(x) < 0$.

The graphs corresponding to cases (a)-(d) are given in Fig. 8.60.

Let us turn our attention to the case (a) (Fig. 8.61). We have $f'(x) > 0$ and $f''(x) > 0$ on $[a, b]$. Draw the secant through the points $A(a, f(a))$ and $B(b, f(b))$. The equation of the secant is

$$\frac{y - f(a)}{f(b) - f(a)} = \frac{x - a}{b - a}.$$



(a) (b)

(c) (d)

Fig. 8.60

The point a_1 at which the secant AB intersects the x-axis lies between a and ξ, being a better approximation to ξ than a. Substituting $y = 0$ into the equation of the secant AB, we easily obtain

$$a_1 = a - \frac{f(a)(b - a)}{f(b) - f(a)}.$$

Look at Fig. 8.59. It is easy to observe that at a_1 the signs of $f(x)$ and $f''(x)$ are opposite.

Now we draw the tangent to the curve $y = f(x)$ at the point $B(b, f(b))$. Notice that at this point $f(x)$ and $f''(x)$ are of the same sign; this is a very important condition: if $f(x)$ and $f''(x)$ are of opposite signs at $B(b, f(b))$

it can happen that the process of approximating the root diverges. The point b_1 of intersection of the tangent with the x-axis lies between b and ξ and offers a better approximation to ξ than b. The tangent is given by the equation

$$y - f(b) = f'(b)(x - b).$$

Then putting $y = 0$, we get

$$b_1 = b - \frac{f(b)}{f'(b)} \quad (f'(b) \neq 0).$$

Therefore we have $a < a_1 < \xi < b_1 < b$.

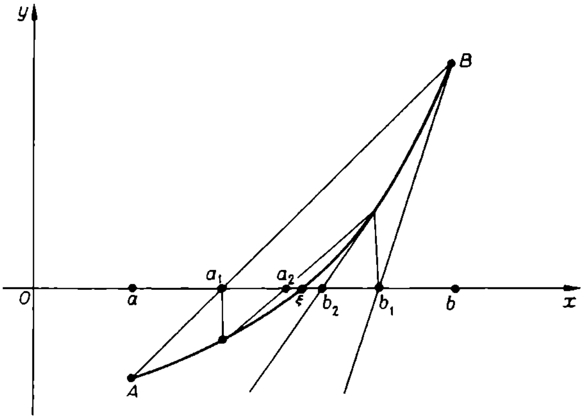


Fig. 8.61

Suppose that the absolute error of the approximation ξ^* to ξ is preassigned. We can estimate the precision of the approximate values a_1 and b_1 by comparing the value $|b_1 - a_1|$ with the preassigned error. If $|b_1 - a_1|$ is greater than the preassigned error then we choose the closed interval $[a_1, b_1]$ and repeat the process by computing the new approximate values for the root ξ as

$$a_2 = a_1 - \frac{f(a_1)(b_1 - a_1)}{f(b_1) - f(a_1)}$$

and

$$b_2 = b_1 - \frac{f(b_1)}{f'(b_1)} \quad (f'(b_1) \neq 0),$$

where $a < a_1 < a_2 < \xi < b_2 < b_1 < b$.

Repeating the process, we finally obtain two sequences of the approximate values for the root

$$a < a_1 < a_2 < \ldots < a_n < \ldots < \xi$$

and

$$b > b_1 > b_2 > \ldots > b_n > \ldots > \xi,$$

where

$$a_n = a_{n-1} - \frac{f(a_{n-1})(b_{n-1} - a_{n-1})}{f(b_{n-1}) - f(a_{n-1})},$$

$$b_n = b_{n-1} - \frac{f(b_{n-1})}{f'(b_{n-1})} \quad (a_0 = a,\ b_0 = b,\ n = 1, 2, 3, \ldots).$$

The sequences $\{a_n\}$ and $\{b_n\}$ are monotone and bounded; hence they have limits. Let

$$\lim_{n\to\infty} a_n = \alpha \quad \text{and} \quad \lim_{n\to\infty} b_n = \beta.$$

It is not hard to verify that $\alpha = \beta = \xi$ where ξ is the unique real root of the equation $f(x) = 0$ provided that the conditions (1)-(3) hold.

Example. Compute the root ξ of the equation $x^2 - 1 = 0$ on the closed interval [0, 2].

◄ Evidently, $\xi = 1$. Let us apply the process outlined above. For the function $f(x) = x^2 - 1$ the following conditions are satisfied:

(1) $f(x)$ is continuous on [0, 2];

(2) $f(0) = -1 < 0$ and $f(2) = 3 > 0$ so that $f(0)\cdot f(2) < 0$;

(3) $f'(x) = 2x$ and $f''(x) = 2$ retain their signs on [0, 2].

Thus there exists the unique root ξ of the equation $x^2 - 1 = 0$ on [0, 2]. Consequently, the process is applicable in this case.

We have $a = 0$ and $b = 2$. For $n = 1$ the recurrence formulas for a_n and b_n give

$$a_1 = 0 - \frac{(-1)\cdot 2}{4} = \frac{1}{2} = 1 - \frac{1}{2},$$

$$b_1 = 2 - \frac{3}{4} = \frac{5}{4} = 1 + \frac{1}{4}.$$

For $n = 2$ we get

$$a_2 = 1 - \frac{1}{14} \quad \text{and} \quad b_2 = 1 + \frac{1}{40},$$

which are the approximate values for ξ with the absolute error $\Delta(\xi^*) < 0.1$. ►

8.13 Taylor's Theorem

Taylor's formula for polynomials. Consider the nth degree polynomial

$$P(x) = b_0 + b_1x + b_2x^2 + \dots + b_nx^n \quad (b_n \neq 0),$$

where $b_0, b_1, b_2, \dots, b_n$ are constant coefficients.

We can express $P(x)$ as the expansion in powers of $x - a$, with some coefficients, where a is an arbitrary number. Indeed, put $x = a + t$. Then $P(x)$ becomes

$$P(x) = P(a + t) = b_0 + b_1(a + t) + \dots + b_n(a + t)^n.$$

On opening the brackets and collecting the similar terms, we can write

$$P(a + t) = A_0 + A_1t + A_2t^2 + A_3t^3 + \dots + A_nt^n$$

or, substituting $x - a$ for t,

$$\begin{aligned} P(x) = {} & A_0 + A_1(x - a) + A_2(x - a)^2 + A_3(x - a)^3 \\ & + \dots + A_n(x - a)^n, \end{aligned} \tag{$*$}$$

where $A_0, A_1, \dots, A_n$ are coefficients dependent on the original coefficients $b_0, b_1, \dots, b_n$.

Differentiating $P(x)$ in $(*)$ n times, we obtain

$$\begin{aligned} P'(x) &= A_1 + 2A_2(x - a) + 3A_3(x - a)^2 + \dots + nA_n(x - a)^{n-1}, \\ P''(x) &= 2 \times 1A_2 + 3 \times 2A_3(x - a) + \dots + n(n - 1)A_n(x - a)^{n-2}, \\ & \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ P^{(n)}(x) &= n(n - 1) \dots 2 \times 1A_n. \end{aligned}$$

Putting $x = a$ in $P(x)$ and in the above identities, we have

$$P(a) = A_0,\ P'(a) = 1! \cdot A_1,\ P''(a) = 2! \cdot A_2,\ \dots,\ P^{(n)}(a) = n! \cdot A_n.$$

Whence

$$A_0 = P(a), \quad A_1 = \frac{P'(a)}{1!}, \quad A_2 = \frac{P''(a)}{2!}, \quad \dots, \quad A_n = \frac{P^{(n)}(a)}{n}.$$

Therefore we can express $P(x)$ as

$$\begin{aligned} P(x) = {} & P(a) + \frac{P'(a)}{1!}(x - a) + \frac{P''(a)}{2!}(x - a)^2 \\ & + \dots + \frac{P^{(n)}(a)}{n!}(x - a)^n. \end{aligned}$$

This is *Taylor's formula*, in powers of $x - a$, for the given polynomial $P(x)$ of degree n or *Taylor's formula* for $P(x)$ at the point a.

If we put $a = 0$ we obtain the specific case of Taylor's formula

$$P(x) = P(0) + \frac{P'(0)}{1!}x + \frac{P''(0)}{2!}x^2 + \dots + \frac{P^{(n)}(0)}{n!}x^n,$$

which is called *Maclaurin's formula.*

Example. Expand the polynomial $P(x) = x^2 - 3x + 2$ in powers of x and in powers of $x - 1$.

◄ Applying Maclaurin's formula, we get

$$\begin{aligned} P(x) &= x^2 - 3x + 2 &&\Rightarrow P(0) = 2, \\ P'(x) &= 2x - 3 &&\Rightarrow P'(0) = -3, \\ P''(x) &= 2 &&\Rightarrow P''(0) = 2, \end{aligned}$$

and

$$P(x) = 2 - \frac{3}{1!}x + \frac{2}{2!}x^2 = 2 - 3x + x^2,$$

so that the expansion of $P(x)$ in powers of x is identical to $P(x)$ itself.

To expand $P(x)$ in powers of $x - 1$ we apply Taylor's formula and get

$$\begin{aligned} P(x) &= x^2 - 3x + 2 &&\Rightarrow P(1) = 0, \\ P'(x) &= 2x - 3 &&\Rightarrow P'(1) = -1, \\ P''(x) &= 2 &&\Rightarrow P''(1) = 2, \end{aligned}$$

and

$$P(x) = 0 - 1(x - 1) + \frac{2}{2!}(x - 1)^2 = -(x - 1) + (x - 1)^2. \quad ►$$

Notice that Taylor's formula gives the value of $P(x)$ at any point x provided that the values of $P(x)$ and all its derivatives at some point a are known.

Taylor's formula for arbitrary functions. Now we consider a function $f(x)$ defined in a neighbourhood of $x = a$; the function $f(x)$ may not be a polynomial of degree $(n - 1)$ but is supposed to have derivatives up to order n in this neighbourhood.

Let us compute the values $f(a), f'(a), \dots, f^{(n-1)}(a)$ and use them to construct the function

$$Q_{n-1}(x) = f(a) + \frac{f'(a)}{1!}(x - a) + \dots + \frac{f^{(n-1)}(a)}{(n - 1)!}(x - a)^{n-1}.$$

Evidently, $Q_{n-1}(x)$ is *Taylor's polynomial* of degree $n - 1$ for the function $f(x)$. If the original function $f(x)$ were a polynomial of degree $(n - 1)$ then we would have the identity $f(x) \equiv Q_{n-1}(x)$ for all x in the neighbourhood in question. However, in the case under consideration the identity does not hold since, by the hypothesis, $f(x)$ is not a polynomial of degree $(n - 1)$.

Let us put

$$f(x) = Q_{n-1}(x) + R_n(x).$$

This equality is called *Taylor's formula* for the function $f(x)$ in the neighbourhood of the point $x = a$ or *Taylor's formula* for the function $f(x)$ at the point a, and $R_n(x)$ is called the nth *remainder* of Taylor's series.

The remainder $R_n(x)$ can be expressed in terms of the nth derivative of the function $f(x)$. For the purpose we assume that $f(x)$ is not a polynomial of degree $(n - 1)$ and possesses continuous derivatives up to order $(n - 1)$ on $[a, b]$ and there exists the nth derivative of $f(x)$ in (a, b). Substituting b for x in Taylor's formula for the function $f(x)$ at the point a, we can write

$$f(b) = f(a) + \frac{f'(a)}{1!}(b - a) + \frac{f''(a)}{2!}(b - a)^2$$

$$+ \ldots + \frac{f^{(n-1)}(a)}{(n - 1)!}(b - a)^{n-1} + R_n.$$

We shall represent R_n in the form

$$R_n = M(b - a)^n,$$

where M is the quantity to be defined.

To this end we consider the auxiliary function

$$\varphi(x) = f(b) - \left[f(x) + \frac{f'(x)}{1!}(b - x) + \frac{f''(x)}{2!}(b - x)^2\right.$$

$$\left. + \ldots + \frac{f^{(n-1)}(x)}{(n - 1)!}(b - x)^{n-1} + M(b - x)^n\right]$$

defined on $[a, b]$. The function $\varphi(x)$ is obtained by replacing a by x in the right-hand side of $f(b)$ and subtracting the right-hand side from $f(b)$.

Observe that $\varphi(x)$ satisfies *Rolle's theorem* (Theorem 8.6). Indeed,

(a) $\varphi(x)$ is continuous on $[a, b]$ since the original function $f(x)$ and all its derivatives up to order $(n - 1)$ are continuous on $[a, b]$;

(b) $\varphi(x)$ has a derivative on (a, b) since the original function $f(x)$ has the nth derivative on (a, b);

(c) $\varphi(x)$ assumes equal values at the endpoints of $[a, b]$: computing $\varphi(x)$ for $x = a$ and $x = b$ gives $\varphi(a) = 0$ and $\varphi(b) = 0$.

Then by virtue of Rolle's theorem there exists a point ξ in (a, b) such that $\varphi'(\xi) = 0$.

Computing the derivative $\varphi'(x)$, we get

$$\varphi'(x) = -\left[f'(x) - \frac{f'(x)}{1!} + \frac{f''(x)}{1!}(b - x) - \frac{f''(x)}{2!}2(b - x) + \right.$$

$$+ \dots + \frac{f^{(n-1)}(x)}{(n-2)!}(b-x)^{n-2} - \frac{f^{(n-1)}(x)}{(n-1)!}(n-1)(b-x)^{n-2}$$

$$+ \frac{f^{(n)}(x)}{(n-1)!}(b-x)^{n-1} - Mn(b-x)^{n-1}\Bigg]$$

so that

$$\varphi'(x) = -\frac{f^{(n)}(x)}{(n-1)!}(b-x)^{n-1} + Mn(b-x)^{n-1}.$$

Therefore

$$\varphi'(\xi) = -(b-\xi)^{n-1}\left[\frac{f^{(n)}(\xi)}{(n-1)!} - Mn\right] = 0.$$

Since $\xi \neq b$ (the point ξ lies inside (a, b)) then

$$\frac{f^{(n)}(\xi)}{(n-1)!} - Mn = 0$$

and

$$M = \frac{f^{(n)}(\xi)}{n!}.$$

Hence we can write $R_n = M(b-a)^n$ as

$$R_n = \frac{f^{(n)}(\xi)}{n!}(b-a)^n \quad (a < \xi < b).$$

Substituting this expression into the formula for $f(b)$, we obtain

$$f(b) = f(a) + \frac{f'(a)}{1!}(b-a) + \frac{f''(a)}{2!}(b-a)^2$$

$$+ \dots + \frac{f^{(n-1)}(a)}{(n-1)!}(b-a)^{n-1} + \frac{f^{(n)}(\xi)}{n!}(b-a)^n.$$

This is Taylor's formula for the function $f(x)$ and

$$R_n = \frac{f^{(n)}(\xi)}{n!}(b-a)^n,$$

where ξ lies between a and b and is the nth *remainder* as given by *Lagrange*.

If we put $n = 1$ Taylor's formula for $f(x)$ becomes as given by Theorem 8.7

$$f(b) = f(a) + \frac{f'(\xi)}{1!}(b-a)$$

or

$$f(b) - f(a) = f'(\xi)(b-a).$$

Taylor's formula remains valid for any points x_0 and x in $[a, b]$ so that we can write it in the form

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2$$
$$+ \dots + \frac{f^{(n-1)}(x_0)}{(n-1)!}(x - x_0)^{n-1} + R_n(x),$$

where $R_n(x) = \dfrac{f^{(n)}(\xi)}{n!}(x - x_0)^n$ and the point ξ lies between x and x_0, or

$$R_n(x) = \frac{f^{(n)}(x_0 + \theta(x - x_0))}{n!}(x - x_0)^n \quad (0 < \theta < 1).$$

Putting $x_0 = 0$, we arrive at *Maclaurin's formula*

$$f(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \dots + \frac{f^{(n-1)}(0)}{(n-1)}x^{n-1} + R_n(x),$$

where $R_n(x) = \dfrac{f^{(n)}(\theta x)}{n!}x^n$, $0 < \theta < 1$.

The remainder in Taylor's formula can also be represented in a form given by Peano. We have assumed that $f(x)$ has the nth derivative $f^{(n)}(x)$ in a neighbourhood of x_0. Now we suppose that $f^{(n)}(x)$ is continuous at x_0. Then $f^{(n)}(x_0 + \theta(x - x_0)) = f^{(n)}(x_0) + \alpha(x)$ where $\alpha(x) \to 0$ as $x \to x_0$ and the nth remainder

$$R_n(x) = \frac{f^{(n)}(x_0 + \theta(x - x_0))}{n!}(x - x_0)^n$$

can be written as

$$R_n(x) = \frac{f^{(n)}(x_0) + \alpha(x)}{n!}(x - x_0)^n$$

or

$$R_n(x) = \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \frac{\alpha(x)(x - x_0)^n}{n!},$$

where $\alpha(x) \to 0$ as $x \to x_0$.

Since $\alpha(x) \to 0$ as $x \to x_0$ then $\dfrac{\alpha(x)(x - x_0)^n}{n!} = o((x - x_0)^n)$ as $x \to x_0$ and the Taylor's formula becomes

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \dots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$
$$+ o((x - x_0)^n) \quad \text{as} \quad x \to x_0.$$

This formula is sometimes called *Taylor's formula for the function $f(x)$ with the remainder as given by Peano.* It is easy to notice that the error of the approximation of $f(x)$ by Taylor's formula is an infinitesimal of higher order than $(x - x_0)^n$ as $x \to x_0$. Hence, this formula is suitable to apply if we wish to approximate $f(x)$ at points sufficiently close to x_0. So it is sometimes called the *local Taylor's formula.*

Maclaurin's formulas for some elementary functions. We shall apply Maclaurin's formula

$$f(x) = f(0) + \frac{f'(0)}{1!} x + \frac{f''(0)}{2!} x^2 + \ldots + \frac{f^{(n-1)}(0)}{(n-1)!} x^{n-1}$$
$$+ \frac{f^{(n)}(\theta x)}{n!} x^n, \quad 0 < \theta < 1$$

to get the approximation of some elementary functions.

Examples. (1) The function $f(x) = e^x$.

◄ We have

$$\begin{array}{ll} f(x) = e^x, & f(0) = 1, \\ f'(x) = e^x, & f'(0) = 1, \\ \ldots\ldots\ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots \\ f^{(n-1)}(x) = e^x, & f^{(n-1)}(0) = 1, \\ f^{(n)}(x) = e^x, & f^{(n)}(\theta x) = e^{\theta x}. \end{array}$$

Applying Maclaurin's formula, we obtain

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots + \frac{x^{n-1}}{(n-1)!} + \frac{e^{\theta x}}{n!} x^n, \quad 0 < \theta < 1.$$

Putting $x = 1$ gives

$$e = 2 + \frac{1}{2!} + \ldots + \frac{1}{(n-1)!} + \frac{e^\theta}{n!}, \quad 0 < \theta < 1.$$

Since

$$0 < \frac{e^\theta}{n!} < \frac{3}{n!}$$

the polynomial $2 + \frac{1}{2!} + \frac{1}{3!} + \ldots + \frac{1}{(n-1)!}$ approximates the number e by defect and with the approximation error less than $\frac{3}{n!}$. ►

(2) The sine function $f(x) = \sin x$.

◄ We have

$$\begin{array}{ll} f(x) = \sin x, & f(0) = 0, \\ f'(x) = \cos x, & f'(0) = 1, \\ f''(x) = -\sin x, & f''(0) = 0, \\ f'''(x) = -\cos x, & f'''(0) = -1, \end{array}$$

and in general

$$f^{(m)}(x) = \sin\left(x + m\frac{\pi}{2}\right).$$

Whence

$$f^{(m)}(0) = \sin m\frac{\pi}{2} = \begin{cases} 0 & \text{for} \quad m = 2k, \\ (-1)^k & \text{for} \quad m = 2k + 1, \end{cases}$$

and

$$f^{(n)}(\theta x) = \sin\left(\theta x + n\frac{\pi}{2}\right).$$

Therefore, the terms with even powers of x vanish so that Taylor's polynomial involving $(2n + 1)$ terms is identically equal to that involving $(2n + 2)$ terms. Then applying Maclaurin's formula and putting $n = 2k + 1$, we obtain

$$\sin x = \frac{x}{1!} - \frac{x^3}{3!} + \frac{x^5}{5!} + \ldots + (-1)^{k-1}\frac{x^{2k-1}}{(2k-1)!} + R_{2k+1}(x),$$

where

$$R_{2k+1}(x) = \frac{x^{2k+1}}{(2k+1)!}\sin\left[\theta x + (2k+1)\frac{\pi}{2}\right], \quad 0 < \theta < 1.$$

Clearly, $|R_{2k+1}(x)| \leqslant \dfrac{|x|^{2k+1}}{(2k+1)!}$. ►

(3) The cosine function $f(x) = \cos x$.

◄ We have

$f(x) = \cos x, \qquad f(0) = 1,$
$f'(x) = -\sin x, \qquad f'(0) = 0,$
$f''(x) = -\cos x, \qquad f''(0) = -1$

and in general

$$f^{(m)}(x) = \cos\left(x + m\frac{\pi}{2}\right)$$

so that

$$f^{(m)}(0) = \cos m\frac{\pi}{2} = \begin{cases} 0 & \text{for} \quad m = 2k + 1, \\ (-1)^k & \text{for} \quad m = 2k. \end{cases}$$

Applying *Maclaurin's formula* for $n = 2k + 2$, we get

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots + (-1)^k\frac{x^{2k}}{(2k)!} + R_{2k+2}(x),$$

where

$$R_{2k+2}(x) = \frac{x^{2k+2}}{(2k+2)!} \cos\left[\theta x + (2k+2)\frac{\pi}{2}\right), \quad 0 < \theta < 1.$$

Evidently, $|R_{2k+2}(x)| \leqslant \frac{x^{2k+2}}{(2k+2)!}$. ►

Maclaurin's formulas for sin x and cos x are suitable when we wish to approximate these functions with any predetermined errors. The Taylor approximations to the functions sin x and cos x in a neighbourhood of $x = 0$ are shown in Fig. 8.62 and in Fig. 8.63.

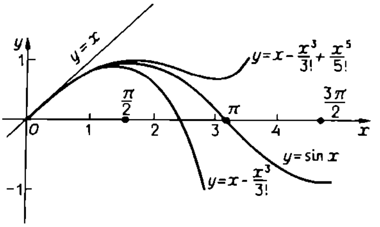


Fig. 8.62

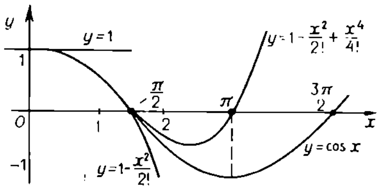


Fig. 8.63

(4) The logarithmic function $f(x) = \ln(1+x)$.

◄ This function is defined and differentiable infinitely many times for $x > -1$.

We have

$$f(x) = \ln(1+x), \qquad f(0) = \ln 1 = 0,$$

$$f'(x) = \frac{1}{1+x}, \qquad f'(0) = 1,$$

$$f''(x) = -\frac{1}{(1/x)^2}, \qquad f''(0) = -1,$$

$$f'''(x) = \frac{1\cdot 2}{(1+x)^3}, \qquad f'''(0) = 2!$$

. . . .

$$f^{(n-1)}(x) = (-1)^n \frac{(n-2)!}{(1+x)^{n-1}}, \qquad f^{(n-1)}(0) = (-1)^n(n-2)!,$$

$$f^{(n)}(x) = (-1)^{n+1}\frac{(n-1)!}{(1+x)^n}, \qquad f^{(n)}(\theta x) = (-1)^{n+1}\frac{(n-1)!}{(1+\theta x)^n}.$$

Using Maclaurin's formula, we obtain

$$\ln(1+x) = \frac{x}{1} - \frac{x^2}{2} + \frac{x^3}{3} - \ldots + (-1)^n \frac{x^{n-1}}{n-1} + R_n(x),$$

where

$$R_n(x) = (-1)^{n+1} \frac{1}{n(1+\theta x)^n} x^n, \quad 0 < \theta < 1. \ ►$$

(5) The power function $f(x) = (1+x)^\alpha$ (α is real, $x > -1$).

◄ We have

$$\begin{aligned} f(x) &= (1+x)^\alpha, & f(0) &= 1; \\ f'(x) &= \alpha(1+x)^{\alpha-1}, & f'(0) &= \alpha; \\ f''(x) &= \alpha(\alpha-1)(1+x)^{\alpha-2}, & f''(0) &= \alpha(\alpha-1); \end{aligned}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\begin{aligned} f^{(n-1)}(x) &= \alpha(\alpha-1)\ldots(\alpha-n+2)(1+x)^{\alpha-n+1}, \\ f^{(n-1)}(0) &= \alpha(\alpha-1)\ldots(\alpha-n+2); \\ f^{(n)}(x) &= \alpha(\alpha-1)\ldots(\alpha-n+1)(1+x)^{\alpha-n}, \\ f^{(n)}(\theta x) &= \alpha(\alpha-1)\ldots(\alpha-n+1)(1+\theta x)^{\alpha-n}. \end{aligned}$$

Whence

$$(1+x)^\alpha = 1 + \frac{\alpha}{1!}x + \frac{\alpha(\alpha-1)}{2!}x^2 + \ldots + \frac{\alpha(\alpha-1)\ldots(\alpha-n+2)}{(n-1)!}x^{n-1} + R_n(x),$$

where

$$R_n(x) = \frac{\alpha(\alpha-1)\ldots(\alpha-n+1)}{n!}(1+\theta x)^{\alpha-n}x^n, \quad 0 < \theta < 1. \ ►$$

It is easy to observe that for natural $\alpha = m$ Maclaurin's formula for $(1+x)^\alpha$ becomes the *binomial formula*

$$(1+x)^m = 1 + \frac{m}{1!}x + \frac{m(m-1)}{2!}x^2 + \ldots + x^m \quad \forall x.$$

Maclaurin's formula and equivalent infinitesimals. We have established equivalence relations for infinitesimals in Chap. 7. These formulas are suitable to represent elementary functions for sufficiently small $|x|$ and to evaluate limits of these functions. However, it is not a rare case when the limit in question depends on higher degrees of x than those involved in equivalence relations previously obtained. In this case we need to extend these equivalence relations to terms involving higher degrees of x to be able to compute the limits of functions as $x \to 0$. These extensions can easily be obtained by applying Maclaurin's formula with remainder specified by Peano. The following is the list of the equivalence relations thus obtained:

$$e^x = 1 + x + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} + o(x^n), \quad x \to 0;$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + o(x^{2n+2}), \quad x \to 0;$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \ldots + (-1)^n \frac{x^{2n}}{(2n)!} + o(x^{2n+1}), \quad x \to 0;$$

$$\ln(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots + (-1)^{n-1} \frac{x^n}{n} + o(x^n), \quad x \to 0;$$

$$(1 + x)^\alpha = 1 + \alpha x + \frac{\alpha(\alpha - 1)}{2!} x^2$$

$$+ \ldots + \frac{\alpha(\alpha - 1) \ldots (\alpha - n + 1)}{n!} x^n + o(x^n), \quad x \to 0.$$

Example. Evaluate $\lim\limits_{x \to 0} \frac{x - \sin x}{x^3}$.

◄ The equivalence relations given in Chap. 7 are of no value here since the principle term of the limit depends on x^3. However using the equivalence relation for $\sin x$ given in the above list, we easily obtain

$$\lim_{x \to 0} \frac{x - \sin x}{x^3} = \lim_{x \to 0} \frac{x - \left(x - \frac{x^3}{3!} + o(x^4)\right)}{x^3}$$

$$= \lim_{x \to 0} \left(\frac{1}{3!} + \frac{o(x^4)}{x^3}\right) = \frac{1}{6}. \quad ►$$

Taylor's theorem and investigating a function for extremum. Taylor's formula is helpful to find extrema of a function.

Theorem 8.21. *Let a function $f(x)$ have the nth derivative $f^{(n)}(x)$ in a neighbourhood of a point x_0. Let $f^{(n)}(x)$ be continuous at x_0 and let $f'(x_0) = f''(x_0) = \ldots = f^{(n-1)}(x_0) = 0$ and $f^{(n)}(x_0) \neq 0$. Then* (i) *if n is odd $f(x)$ has no extremum at x_0;* (ii) *if n is even and $f^{(n)}(x_0) < 0$ then $f(x)$ has a maximum at x_0 and* (iii) *if n is even and $f^{(n)}(x_0) > 0$ $f(x)$ has a minimum at x_0.*

◄ Observe that $f(x)$ has an extremum at x_0 if there exists $\delta > 0$ such that the difference $f(x) - f(x_0)$ retains its sign on the interval $(x_0 - \delta, x_0 + \delta)$.

Applying Taylor's formula, we have

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2$$

$$+ \dots + \frac{f^{(n-1)}(x_0)}{(n-1)!}(x - x_0)^{n-1} + \frac{f^{(n)}(x_0 + \theta(x - x_0))}{n!}(x - x_0)^n,$$

where $0 < \theta < 1$. Whence

$$f(x) - f(x_0) = \frac{f^{(n)}(x_0 + \theta(x - x_0))}{n!}(x - x_0)^n \qquad (*)$$

since $f'(x_0) = f''(x_0) = \dots = f^{(n-1)}(x_0) = 0$.

By the hypothesis $f^{(n)}(x)$ is continuous at x_0 and $f^{(n)}(x_0) \neq 0$. This implies (see Theorem 7.22) that there exists $\delta > 0$ such that at every point of $(x_0 - \delta, x_0 + \delta)$ $f^{(n)}(x)$ is of the same sign as $f^{(n)}(x_0)$.

The following cases have to be distinguished:

(a) n is even and $f^{(n)}(x_0) > 0$. Then

$$\frac{f^{(n)}(x_0 + \theta(x - x_0))}{n!}(x - x_0)^n \geqslant 0 \quad \forall x \in (x_0 - \delta, x_0 + \delta)$$

so that $(*)$ gives

$$f(x) - f(x_0) \geqslant 0 \quad \forall x \in (x_0 - \delta, x_0 + \delta).$$

By the definition of a maximum of a function we infer that $f(x)$ has a minimum at x_0.

(b) n is even and $f^{(n)}(x_0) < 0$. Then we have

$$\frac{f^{(n)}(x_0 + \theta(x - x_0))}{n!}(x - x_0)^n \leqslant 0 \quad \forall x \in (x_0 - \delta, x_0 + \delta)$$

and $f(x) - f(x_0) \leqslant 0$ for all $x \in (x_0 - \delta, x_0 + \delta)$ so that $f(x)$ has a maximum at x_0.

(c) n is odd and $f^{(n)}(x_0) \neq 0$. Then

$$\frac{f^{(n)}(x_0 + \theta(x - x_0))}{n!}(x - x_0)^n \quad \text{and} \quad f^{(n)}(x_0)$$

are of the same sign whenever $x > x_0$ and of opposite signs as $x < x_0$. Hence, for any whatever small $\delta > 0$ the difference $f(x) - f(x_0)$ changes its sign on the interval $(x_0 - \delta, x_0 + \delta)$. This means that $f(x)$ has no extremum at x_0. ►

Example. Investigate the functions $y = x^4$ and $y = x^3$ for extremum.

◄ It is easy to see that the point $x = 0$ is a critical point for these functions. The first three derivatives of $y = x^4$ are zeros at $x = 0$ and $f^{(4)}(0) = 24 > 0$. Hence, $n = 4$ is even and $f^{(4)}(0) > 0$ so that $y = x^4$ has a minimum at $x = 0$.

The function $y = x^3$ has no extremum at $x = 0$ since its first two derivatives are zeros at $x = 0$ and the third derivative is not equal to zero so that $n = 3$ is odd. ►

Remark. Taylor's formula is helpful to prove the following theorem which specifies the sufficient conditions for a function to have a point of inflection at x_0.

Theorem 8.22. *Let $f(x)$ have the nth derivative $f^{(n)}(x)$ in a neighbourhood of x_0 and let $f^{(n)}(x)$ be continuous at x_0. Moreover, let $f''(x_0) = f'''(x_0) = \ldots = f^{(n-1)}(x_0) = 0$ and $f^{(n)}(x_0) \neq 0$. Then $M_0(x_0, f(x_0))$ is the point of inflection of the graph of $y = f(x)$ if n is odd.*

To verify this theorem it suffices to consider the function $f(x) = x^3$.

8.14 Vector Function of a Scalar Argument

Definition. Suppose a point M moves along some trajectory L. We can locate this point at any time t by indicating its position vector $\mathbf{r}$, its velocity vector $\mathbf{v}$, its acceleration vector $\mathbf{w}$, etc. Each of these vectors can be regarded as a vector function of a scalar argument t, i.e., $\mathbf{r} = \mathbf{r}(t)$, $\mathbf{v} = \mathbf{v}(t)$, $\mathbf{w} = \mathbf{w}(t)$.

We say that a vector function of a scalar argument t is defined on the interval (α, β) and write $\mathbf{a} = \mathbf{a}(t)$ if there is a certain law which assigns a definite vector $\mathbf{a}$ to each t from (α, β).

Let $\mathbf{a}$ be expanded relative to the unit vectors $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ of some system of coordinates so that

$$\mathbf{a} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}.$$

If $\mathbf{a} = \mathbf{a}(t)$ is a vector function of the argument t then the coordinates x, y, z of $\mathbf{a} = \mathbf{a}(t)$ become scalar functions of the argument t, i.e.,

$$x = \varphi(t),\ y = \psi(t),\ z = \gamma(t),\ \alpha < t < \beta.$$

Conversely, if the coordinates x, y, z of the vector $\mathbf{a}$ are functions of the argument t so is the vector $\mathbf{a}$, i.e.,

$$\mathbf{a} = \varphi(t)\mathbf{i} + \psi(t)\mathbf{j} + \gamma(t)\mathbf{k}.$$

Therefore the vector function $\mathbf{a}(t)$ is fully determined by the scalar functions $x = \varphi(t)$, $y = \psi(t)$ and $z = \gamma(t)$ and vice versa.

The length and direction of the vector $\mathbf{a}(t)$ can in general be different for different t; also, $\mathbf{a}(t)$ can be applied at different points for different t, for example as the velocity vector of a moving particle.

Suppose now that the vector $\mathbf{a}(t)$ starts at a given point O in space. For different t $\mathbf{a}(t)$ ends at different points which form some set of points in this space. This set of points corresponding to all possible terminal points of $\mathbf{a}(t)$ when t varies over its domain of definition is called the ***hodograph*** of the vector function $\mathbf{a}(t)$. In general the hodograph of $\mathbf{a}(t)$ is a curve which contains all points of the space where $\mathbf{a}(t)$ ends (Fig. 8.64). Observe that the hodograph of the position vector $\mathbf{r}$ of a moving point coincides with the trajectory L of this point.

The equation

$$\mathbf{r} = \mathbf{r}(t), \quad \alpha < t < \beta$$

or

$$\mathbf{r}(t) = \varphi(t)\mathbf{i} + \psi(t)\mathbf{j} + \gamma(t)\mathbf{k}$$

is called the ***vector equation*** of the curve L.

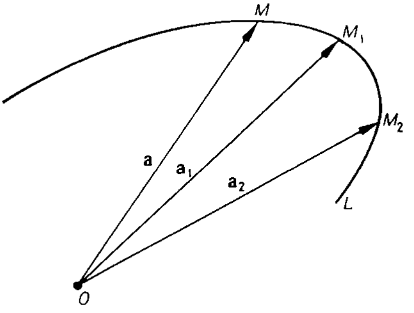


Fig. 8.64

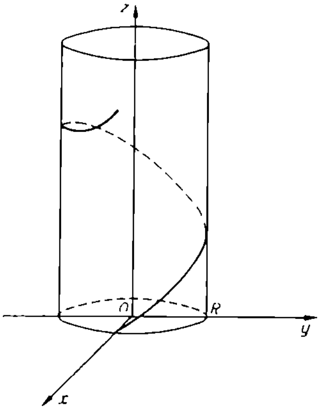


Fig. 8.65

The equations

$$\begin{cases} x = \varphi(t), \\ y = \psi(t), \\ z = \gamma(t), \end{cases} \qquad \alpha < t < \beta$$

are called the *parametric equations of the curve L*.

For example, the equations

$$\begin{cases} x = R\cos t, \\ y = R\sin t, \\ z = ht, \end{cases} \qquad 0 \leqslant t < 2\pi \quad (R,\ h \text{ are constant})$$

are the parametric equations which describe one turn of a *screw line* (or a *circular helix*) in space (Fig. 8.65).

Limits and continuity of vector functions. Let $\mathbf{a} = \mathbf{a}(t)$ be defined in a neighbourhood of a point $t = t_0$ except probably at $t = t_0$. The constant vector $\mathbf{A}$ is called the *limit* of the vector function $\mathbf{a}(t)$ as $t \to t_0$ if, given any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|\mathbf{a}(t) - \mathbf{A}| < \varepsilon$$

whenever $|t - t_0| < \delta$ and $t \neq t_0$.

In symbols one writes

$$\lim_{t \to t_0} \mathbf{a}(t) = \mathbf{A}.$$

The geometric interpretation of the notion of a limit of a vector function $\mathbf{a}(t)$ is that the length of the vector $\mathbf{a}(t) - \mathbf{A}$ tends to zero as $t \to t_0$ so that the vector $\mathbf{a}(t)$ tends to coincide with the vector $\mathbf{A}$ as $t \to t_0$ (Fig. 8.66). Hence

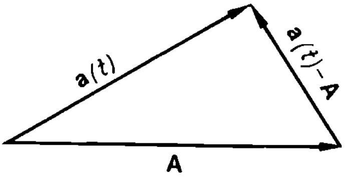


Fig. 8.66

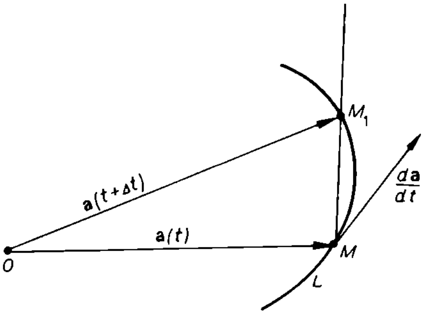


Fig. 8.67

$$(\lim_{t \to t_0} \mathbf{a}(t) = \mathbf{A}) \Leftrightarrow (\lim_{t \to t_0} |\mathbf{a}(t) - \mathbf{A}| = 0).$$

Let $\mathbf{a}(t) = \varphi(t)\mathbf{i} + \psi(t)\mathbf{j} + \gamma(t)\mathbf{k}$ and $\mathbf{A} = a\mathbf{i} + b\mathbf{j} + c\mathbf{k}$. Then

$$|\mathbf{a}(t) - \mathbf{A}| = \sqrt{(\varphi(t) - a)^2 + (\psi(t) - b)^2 + (\gamma(t) - c)^2}.$$

Whence, if $\lim_{t \to t_0} \mathbf{a}(t) = \mathbf{A}$ then

$$\lim_{t \to t_0} \varphi(t) = a, \quad \lim_{t \to t_0} \psi(t) = b, \quad \lim_{t \to t_0} \gamma(t) = c$$

and vice versa.

Let a vector function $\mathbf{a}(t)$ be defined on the interval $\alpha < t < \beta$ and $t_0 \in (\alpha, \beta)$. We say that the vector function $\mathbf{a}(t)$ is *continuous* at $t = t_0$ if

$$\lim_{t \to t_0} \mathbf{a}(t) = \mathbf{a}(t_0).$$

Differentiation of vector functions. Let $\mathbf{a} = \mathbf{a}(t)$ be a vector function defined on the interval $\alpha < t < \beta$ and let a curve L be a hodograph of $\mathbf{a}(t)$. Then a certain value $t \in (\alpha, \beta)$ corresponds to the point M on L. Suppose that Δt is an arbitrary increment in t such that $t + \Delta t \in (\alpha, \beta)$. Then the vector $\mathbf{a}(t + \Delta t)$ gives the point M_1 on L (Fig. 8.67).

Consider the increment $\Delta \mathbf{a}$ in the vector function $\mathbf{a}(t)$ corresponding to the increment Δt so that

$$\Delta \mathbf{a} = \mathbf{a}(t + \Delta t) - \mathbf{a}(t).$$

The ratio

$$\frac{\Delta \mathbf{a}}{\Delta t} = \frac{\mathbf{a}(t + \Delta t) - \mathbf{a}(t)}{\Delta t} \quad (\Delta t \neq 0)$$

is a vector collinear to the vector $\Delta \mathbf{a}$.

Let the ratio $\frac{\Delta \mathbf{a}}{\Delta t}$ have a limit as $\Delta t \to 0$. Then this limit is called the *derivative* of the vector function $\mathbf{a}(t)$ with respect to the scalar argument t at the point t, customarily denoted by $\frac{d\mathbf{a}(t)}{dt}$ or by $\mathbf{a}'(t)$. Hence

$$\frac{d\mathbf{a}(t)}{dt} = \lim_{\Delta t \to 0} \frac{\Delta \mathbf{a}}{\Delta t} = \lim_{\Delta t \to 0} \frac{\mathbf{a}(t + \Delta t) - \mathbf{a}(t)}{\Delta t}.$$

The vector function $\mathbf{a}(t)$ having a derivative at a point t is called *differentiable* at t.

Now we shall investigate how the vector $\frac{d\mathbf{a}}{dt}$ is directed. The point M_1 moves along the hodograph L towards the point M as $\Delta t \to 0$ so that the secant MM_1 tends to the tangent to L at M. Hence, the derivative $\frac{d\mathbf{a}}{dt}$ is a vector tangent to the hodograph of $\mathbf{a}(t)$ at M and the direction of $\frac{d\mathbf{a}}{dt}$ coincides with that of the motion of the terminal point of the vector $\mathbf{a}(t)$ caused by the increase of the parameter t.

The derivative $\frac{d\mathbf{a}}{dt}$ can be expressed in terms of the coordinates of $\mathbf{a}(t)$. Let $\mathbf{a}(t) = \varphi(t)\mathbf{i} + \psi(t)\mathbf{j} + \gamma(t)\mathbf{k}$. Then

$$\Delta \mathbf{a}(t) = \mathbf{a}(t + \Delta t) - \mathbf{a}(t) = \mathbf{i}\Delta\varphi(t) + \mathbf{j}\Delta\psi(t) + \mathbf{k}\Delta\gamma(t).$$

Dividing both sides by $\Delta t \neq 0$, we obtain

$$\frac{\Delta \mathbf{a}(t)}{\Delta t} = \mathbf{i}\frac{\Delta\varphi(t)}{\Delta t} + \mathbf{j}\frac{\Delta\psi(t)}{\Delta t} + \mathbf{k}\frac{\Delta\gamma(t)}{\Delta t}.$$

If the functions $\varphi(t)$, $\psi(t)$ and $\gamma(t)$ have derivatives at a given point t then every summand on the right has a limit as $\Delta t \to 0$ so that there exists a limit of the left-hand side of the above expression, i.e., there exists the derivative $\frac{d\mathbf{a}(t)}{dt}$. Evaluating the limit as $\Delta t \to 0$, we get

$$\frac{d\mathbf{a}}{dt} = \mathbf{i}\frac{d\varphi}{dt} + \mathbf{j}\frac{d\psi}{dt} + \mathbf{k}\frac{d\gamma}{dt}. \qquad (*)$$

Hence, relative to a stationary system of coordinates the derivative $\frac{d\mathbf{a}}{dt}$ of the vector $\mathbf{a}(t)$ is given by the formula (*) so that in order to compute the derivative of the vector function $\mathbf{a}(t)$ we have to compute the derivatives of the coordinates of $\mathbf{a}(t)$.

Let $\mathbf{r} = \mathbf{r}(t)$ be a position vector of a point moving in space. Then the velocity of this point at the moment t is determined by the derivative of $\mathbf{r}(t)$ at t:

$$\frac{d\mathbf{r}(t)}{dt} = \mathbf{v}(t),$$

Example. Find the derivative of the vector function

$$\mathbf{a}(t) = \mathbf{i}R\cos t + \mathbf{j}R\sin t + \mathbf{k}ht \quad (R,\ h \text{ are constant}).$$

◄ Applying (*), we get

$$\frac{d\mathbf{a}}{dt} = -\mathbf{i}R\sin t + \mathbf{j}R\cos t + \mathbf{k}h. \quad ►$$

Differentiation laws. (1) If $\mathbf{c}$ is a constant vector then $\frac{d\mathbf{c}}{dt} = 0$.

(2) If $\mathbf{a}(t)$ and $\mathbf{b}(t)$ have derivatives at t then

$$\frac{d}{dt}(\mathbf{a}(t) \pm \mathbf{b}(t)) = \frac{d\mathbf{a}(t)}{dt} \pm \frac{d\mathbf{b}(t)}{dt}.$$

(3) If a vector function is multiplied by a scalar its derivative is multiplied by this scalar so that

$$\frac{d(\alpha\mathbf{a}(t))}{dt} = \alpha\frac{d\mathbf{a}(t)}{dt} \quad (\alpha \text{ is a constant}).$$

(4) The derivative of the scalar product of vectors is given by

$$\frac{d}{dt}(\mathbf{a}(t), \mathbf{b}(t)) = \left(\frac{d\mathbf{a}}{dt}, \mathbf{b}\right) + \left(\mathbf{a}, \frac{d\mathbf{b}}{dt}\right).$$

Corollary. If $\mathbf{a}(t)$ is a unit vector, i.e., $|\mathbf{e}(t)| = 1$, then $\frac{d\mathbf{e}}{dt}$ is perpendicular to $\mathbf{e}$.

◄ Indeed, if **e** is a unit vector then (**e**, **e**) = 1. On differentiating (**e**, **e**) = 1, we have

$$\left(\frac{d\mathbf{e}}{dt}, \mathbf{e}\right) + \left(\mathbf{e}, \frac{d\mathbf{e}}{dt}\right) = 0$$

or

$$2\left(\frac{d\mathbf{e}}{dt}, \mathbf{e}\right) = 0.$$

Whence we infer that $\frac{d\mathbf{e}}{dt}$ is perpendicular to **e**. ►

(5) The derivative of the vector product of the vectors is given by

$$\frac{d}{dt}[\mathbf{a}(t), \mathbf{b}(t)] = \left[\frac{d\mathbf{a}}{dt}, \mathbf{b}\right] + \left[\mathbf{a}, \frac{d\mathbf{b}}{dt}\right].$$

Exercises

Compute the derivatives of the following functions

1. $y = x^2 - 5x + 1$, **2.** $y = 2\sqrt{x} - \frac{1}{x} + \sqrt[3]{3}$.

3. $y = (x^2 - 3x + 3)(x^3 - 1)$. **4.** $y = (\sqrt{x} + 1)\left(\frac{1}{\sqrt{x}} - 1\right)$. **5.** $y = \frac{x}{x^2 + 1}$.

6. $y = (x^3 + 1)\left(5 - \frac{1}{x^2}\right)$, $y'(1) = ?$ **7.** $y = \sqrt[3]{\frac{1}{1 + x^2}}$.

8. $y = \sin x - \cos x$. **9.** $y = \frac{\sin x}{1 + \cos x}$. **10.** $y = \sin^2 x$.

11. $y = \frac{1}{3}\tan^3 x - \tan x + x$. **12.** $y = \sin 5x$. **13.** $y = 2\sin(3x - 1)$.

14. $y = \sin 1/x$. **15.** $y = \sin^5 2x$. **16.** $y = \sin(\sin x)$. **17.** $y = x\sin^{-1} x$.

18. $y = x\sin x\tan^{-1} x$. **19.** $y = \sin^{-1}\frac{2}{x}$. **20.** $y = \tan^{-1} x^2$.

21. $y = \ln^2 x$. **22.** $y = x^2\log_3 x$. **23.** $y = \frac{\ln x}{1 + x^2}$. **24.** $y = \ln\tan x$.

25. $y = \frac{1}{\ln x}$. **26.** $y = 9^x$. **27.** $y = x e^x$. **28.** $y = \frac{x^3 + 2^x}{x}$.

29. $y = x^3 - 3^x$. **30.** $y = 10^{3x+1}$. **31.** $y = 5^{\sin x}$. **32.** $y = \sin(3^x)$.

33. $y = \sinh^3 x$. **34.** $y = \sqrt{\cosh x}$. **35.** $y = \tanh(\ln x)$. **36.** $y = 3^{\sinh^2 x}$.

37. $y = (\sin x)^{\cos x}$. **38.** $y = x^{\sin x}$. **39.** $y = x^{\ln x}$.

40. $y = \dfrac{(x-2)^2 \sqrt[3]{x+1}}{(x-5)^3}$. **41.** $y = \dfrac{x}{2}\sqrt{x^2 - a^2} - \dfrac{a^2}{2} \ln(x + \sqrt{x^2 - a^2})$.

42. $y = \dfrac{1}{2} \ln \tan \dfrac{x}{2} - \dfrac{1}{2} \dfrac{\cos x}{\sin^2 x}$.

43. $y = 3x^3 \sin^{-1} x + (x^2 + 2)\sqrt{1 - x^2}$.

44. $y = x(\sin^{-1} x)^2 - 2x + 2\sqrt{1 - x^2} \sin^{-1} x$.

Compute the first differentials of the functions

45. $y = \dfrac{1}{4x^4}$. **46.** $y = \tan^2 x$. **47.** $y = 5^{\ln \sin x}$.

48. Compute the approximate value of $\tan^{-1} 1.02$.

Compute the derivatives of higher orders for the functions

49. $y = x^3 - 3x + 5$, $y'' = ?$ **50.** $y = \tan^{-1} x$, $y''(1) = ?$

51. $y = a^{3x}$, $y''' = ?$ **52.** $y = x^5 \ln x$, $y''' = ?$

Write down the nth derivatives of the following functions

53. $y = x\,e^x$. **54.** $y = \sin^2 x$. **55.** $y = \dfrac{1}{x(1+x)}$.

Find the differentials of higher orders for the functions

56. $y = e^{-x^2/2}$, $d^2y = ?$ **57.** $y = x^m$, $d^3y = ?$

Compute the derivatives $\dfrac{dy}{dx}$ of the functions given parametrically

58. $\begin{cases} x = 2t - 1, \\ y = t^3. \end{cases}$ **59.** $\begin{cases} x = a\cos^3 t, \\ y = b\sin^3 t. \end{cases}$ **60.** $\begin{cases} x = e^{-t}, \\ y = e^{2t}. \end{cases}$

Compute the second derivatives $\dfrac{d^2y}{dx^2}$ of the functions given parametrically

61. $\begin{cases} x = \ln t, \\ y = t^3. \end{cases}$ **62.** $\begin{cases} x = e^t \cos t, \\ y = e^t \sin t. \end{cases}$

Draw the graphs of the following functions

63. $y = x - x^3$. **64.** $y = \dfrac{x^2 - 1}{x}$. **65.** $y = \dfrac{x^2 + 1}{x}$. **66.** $y = \dfrac{x^3 - 4}{x^2}$.

67. $y = xe^{-x}$. **68.** $y = \dfrac{e^{2(x+1)}}{2(x+1)}$. **69.** $y = 2\ln \dfrac{x-1}{x} + 1$.

70. $y = \dfrac{-8 - x^2}{\sqrt{x^2 - 4}}$. **71.** $y = \sqrt[3]{x(x-1)^2}$. **72.** $y = \sqrt[3]{x(x-2)}$.

73. $y = \sqrt[3]{(x-1)^2} - \sqrt[3]{(x-2)^2}$.

Compute the maximum and minimum for each of the following functions defined on closed intervals

74. $y = 4 - x - \dfrac{4}{x^2}$ for $x \in [1, 4]$.

75. $y = \sqrt[3]{2x^2(x-6)}$ for $x \in [-2, 4]$.

76. $y = \dfrac{10x + 10}{x^2 + 2x + 2}$ for $x \in [-1, 2]$.

77. Expand the polynomial $x^3 + 3x^2 - 2x + 4$ in powers of $x + 1$.

78. Expand the polynomial $x^3 - 2x^2 + 3x + 5$ in powers of $x - 2$.

79. Use Taylor's formula to expand the function $f(x) = 1/x$ in a neighbourhood of the point $x_0 = -1$.

80. Use Maclaurin's formula to expand the function $f(x) = x\,\mathrm{e}^x$ in a neighbourhood of the point $x_0 = 0$.

81. Use Maclaurin's formula to find the expansion of the function $f(x) = \mathrm{e}^{x/2+2}$ with the remainder $o(x^n)$.

82. The shape of a string hanging of its own weight alone is given by the equation of the catenary $y = a \cosh x/a$ where a is a constant. Show that given sufficiently small $|x|$ the string takes the shape of a parabola.

Use the derivatives of higher orders to investigate the shape for each of the following functions in a neighbourhood of a given point x_0.

83. $y = \sin^2(x-1) - x^2 + 2x$, $x_0 = 1$.

84. $y = \cos x + \cosh x$, $x_0 = 0$.

85. $y = x^2 + 2\ln(x+2)$, $x_0 = -1$.

86. $y = x^2 - 2\mathrm{e}^{x-1}$, $x_0 = 1$.

Answers

1. $y' = 2x - 5$. **2.** $y' = \dfrac{1}{\sqrt{x}} + \dfrac{1}{x^2}$. **3.** $y' = 5x^4 - 12x^3 + 9x^2 - 2x + 3$.

4. $y' = -\dfrac{1}{2\sqrt{x}}\left(1 + \dfrac{1}{x}\right)$. **5.** $y' = \dfrac{1 - x^2}{(x^2+1)^2}$. **6.** $y'(1) = 16$. **7.** $y' = -\dfrac{2x}{3(1+x^2)^{4/3}}$.

8. $y' = \cos x + \sin x$. **9.** $y' = \dfrac{1}{1 + \cos x}$. **10.** $y' = \sin 2x$. **11.** $y' = \tan^4 x$.

12. $y' = 5\cos 5x$. **13.** $y' = 6\cos(3x - 1)$. **14.** $y' = -\dfrac{1}{x^2}\cos\dfrac{1}{x}$.

15. $y' = 10\sin^4 2x\cos 2x$. **16.** $y' = \cos(\sin x)\cos x$. **17.** $y' = \sin^{-1}x + \dfrac{x}{\sqrt{1-x^2}}$.

18. $y' = \sin x\tan^{-1}x + x\cos x\tan^{-1}x + \dfrac{x\sin x}{1+x^2}$. **19.** $y' = -\dfrac{2}{|x|\sqrt{x^2-4}}$.

20. $y' = \dfrac{2x}{1+x^4}$. **21.** $y' = \dfrac{2}{x}\ln x$. **22.** $y' = 2x\log_3 x + \dfrac{x}{\ln 3}$.

23. $y' = \dfrac{1+x^2-2x^2\ln x}{x(1+x^2)^2}$. **24.** $y' = \dfrac{2}{\sin 2x}$. **25.** $y' = -\dfrac{1}{x\ln^2 x}$. **26.** $y' = 9^x\ln 9$.

27. $y' = e^x(x+1)$. **28.** $y' = \dfrac{2^x(x\ln 2 - 1) + 2x^3}{x^2}$. **29.** $y' = 3x^2 - 3^x\ln 3$.

30. $y' = 3\cdot 10^{3x+1}\ln 10$. **31.** $y' = 5^{\sin x}\cos x\ln 5$. **32.** $y' = 3^x\cos(3x)\ln 3$.

33. $y' = 3\sinh^2 x\cosh x$. **34.** $y' = \dfrac{\sinh x}{2\sqrt{\cosh x}}$. **35.** $y' = \dfrac{1}{x\cosh^2(\ln x)}$.

36. $y' = 3^{\sinh^2 x}\sinh 2x\ln 3$. **37.** $y' = (\sin x)^{\cos x}\left(\dfrac{\cos^2 x}{\sin x} - \sin x\ln(\sin x)\right)$.

38. $y' = x^{\sin x}\left(\cos(\ln x) + \dfrac{\sin x}{x}\right)$. **39.** $y' = 2x^{\ln x - 1}\ln x$.

40. $y' = \dfrac{2(x-2)(x^2+11x+1)}{3(x-5)^4\sqrt[3]{(x+1)^2}}$. **41.** $y' = \sqrt{x^2-a^2}$. **42.** $y' = \dfrac{1}{\sin^3 x}$.

43. $y' = 9x^2\sin^{-1}x$. **44.** $y' = (\sin^{-1}x)^2$. **45.** $dy = -\dfrac{dx}{x^5}$. **46.** $dy = \dfrac{2\tan x}{\cos^2 x}dx$.

47. $dy = 5^{\ln\sin x}\cot x\ln 5\,dx$. **48.** $\tan^{-1}1.02 \approx 0.795$. **49.** $y'' = 6x$. **50.** $y''(1) = -\dfrac{1}{2}$.

51. $y''' = 27a^{3x}\ln^3 a$. **52.** $y''' = x^2(60\ln x + 47)$. **53.** $y^{(n)} = e^x(x+n)$.

54. $y^{(n)} = 2^{n-1}\sin\left[2x + (n-1)\dfrac{\pi}{2}\right]$. **55.** $y^{(n)} = (-1)^n\cdot n!\left[\dfrac{1}{x^{n+1}} - \dfrac{1}{(1+x)^{n+1}}\right]$

$\left(\textit{Hint.}\text{ Express the function in the form } y = \dfrac{1}{x(1+x)} = \dfrac{1+x-x}{x(1+x)} = \dfrac{1}{x} - \dfrac{1}{1+x}\right)$.

56. $d^2y = e^{-x^2/2}(x^2-1)dx^2$. **57.** $d^3y = m(m-1)(m-2)x^{m-3}dx^3$. **58.** $\dfrac{dy}{dx} = \dfrac{3}{2}t^2$.

59. $\dfrac{dy}{dx} = -\dfrac{b}{a}\tan t$. **60.** $\dfrac{dy}{dx} = -2e^{3t}$. **61.** $\dfrac{d^2y}{dx^2} = 9t^3$. **62.** $\dfrac{d^2x}{dy^2} = \dfrac{2e^{-t}}{(\cos t - \sin t)^3}$.

63. Fig. 8.68. **64.** Fig. 8.69. **65.** Fig. 8.70. **66.** Fig. 8.71. **67.** Fig. 8.72. **68.** Fig. 8.73.
69. Fig. 8.74. **70.** Fig. 8.75. **71.** Fig. 8.76. **72.** Fig. 8.77. **73.** Fig. 8.78. **74.** $M = 1$, $m = -1$.
75. $M = 0$, $m = -4$. **76.** $M = 5$, $m = 0$. **77.** $(x+1)^3 - 5(x+1) + 8$.
78. $(x-2)^3 + 4(x-2)^2 + 7(x-2) + 11$. **79.** $f(x) = -1 - (x+1) - (x+1)^2 - \dots - (x+1)^{n-1} + (-1)^n\dfrac{(x+1)^n}{[-1+\theta(x+1)]^{n+1}}$, $0 < \theta < 1$. **80.** $f(x) = x + \dfrac{x^2}{1!} + \dfrac{x^3}{2!} + \dots + \dfrac{x^{n-1}}{(n-2)!} + \dfrac{x^n}{n!}(\theta x + n)e^{\theta x}$, $0 < \theta < 1$. **81.** $f(x) = \sum\limits_{k=0}^{n}\dfrac{e^2}{2^k k!}x^k + o(x^n)$. **83.** Fig. 8.79.
84. Fig. 8.80. **85.** Fig. 8.81. **86.** Fig. 8.82.

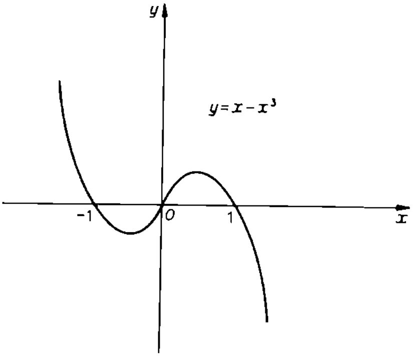


Fig. 8.68

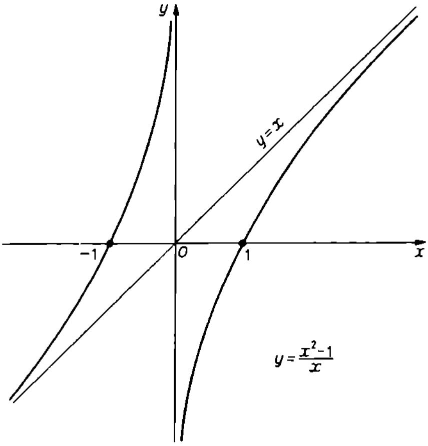


Fig. 8.69

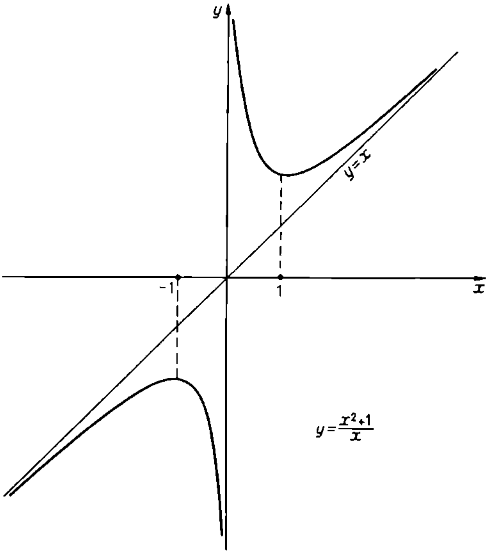


Fig. 8.70

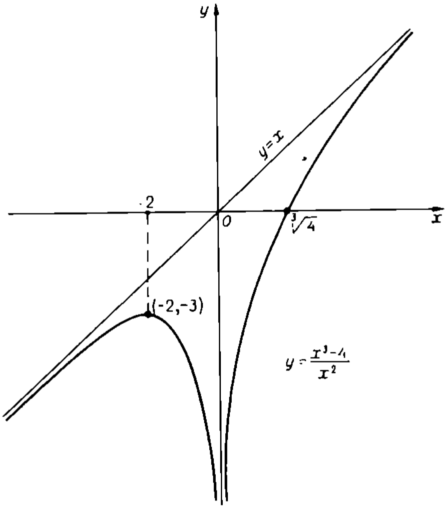


Fig. 8.71

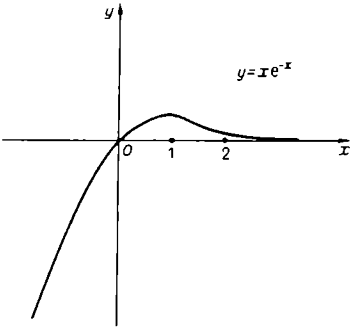


Fig. 8.72

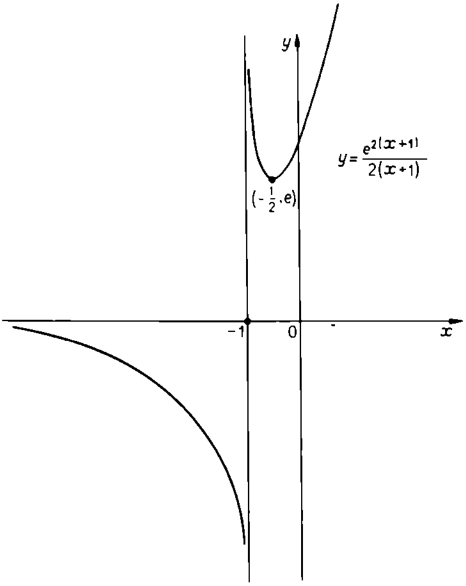


Fig. 8.73

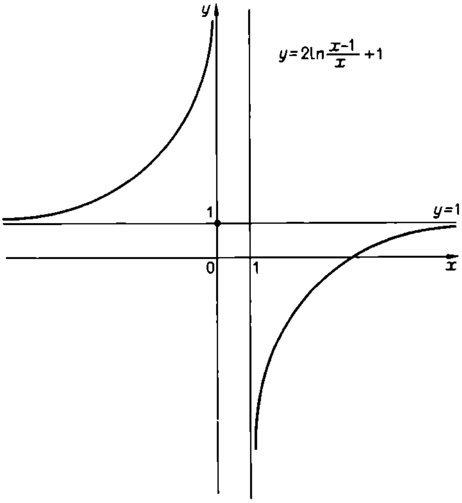


Fig. 8.74

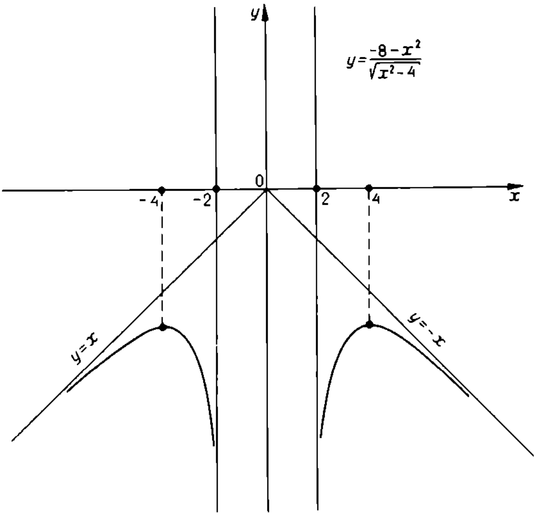


Fig. 8.75

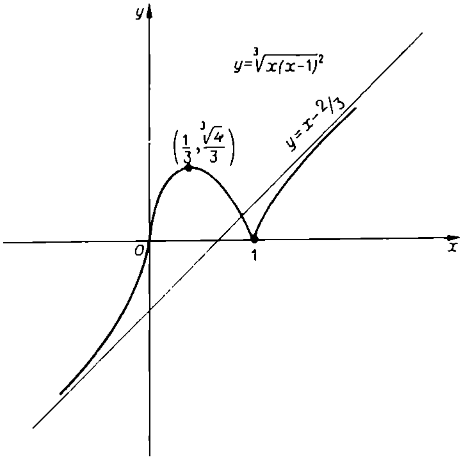


Fig. 8.76

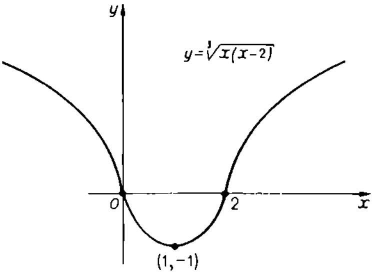


Fig. 8.77

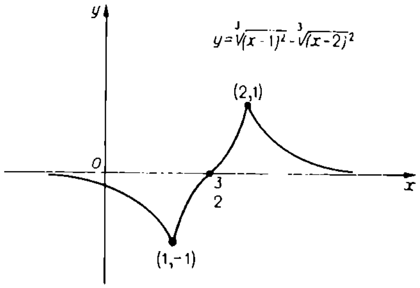


Fig. 8.78

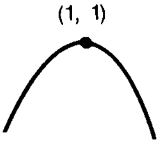


Fig. 8.79

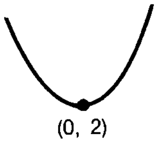


Fig. 8.80

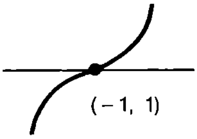


Fig. 8.81

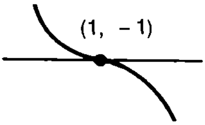


Fig. 8.82

Chapter 9

Integral Calculus. The Indefinite Integral

9.1 Basic Concepts and Definitions

Antiderivative. *Differential calculus* provides methods of computing the derivative $f'(x)$ of a given function $f(x)$. Now we shall examine approaches and methods of solving the problem inverse to that of differentiating functions; namely, we shall study the problem of finding the function $f(x)$ provided that its derivative $f'(x)$ is given. The branch of mathematics that deals with methods and techniques of solving the latter problem is called *integral calculus.*

We say that the function $F(x)$ is the *antiderivative* (or *primitive*) of the function $f(x)$ on the finite or infinite open interval (a, b) if $F(x)$ is differentiable at every point of (a, b) and $F'(x) = f(x)$ or, equivalently, $dF(x) = f(x)\,dx$ for all $x \in (a, b)$.

Examples. (1) The function $F(x) = \sin^{-1}x$ is the antiderivative of the function $f(x) = \dfrac{1}{\sqrt{1 - x^2}}$ on the interval $(-1, 1)$ since for all $x \in (-1, 1)$

$$F'(x) = (\sin^{-1}x)' = \frac{1}{\sqrt{1 - x^2}} .$$

(2) The function $F(x) = \dfrac{a^x}{\ln a}$, $0 < a \neq 1$, is the antiderivative of the function $f(x) = a^x$ on the interval $(-\infty, +\infty)$ since for all $x \in (-\infty, +\infty)$

$$F'(x) = \left(\frac{a^x}{\ln a}\right)' = \frac{a^x \ln a}{\ln a} = a^x.$$

If $F(x)$ is an antiderivative of the function $f(x)$ on the interval (a, b) so is the function $\Phi(x) = F(x) + C$ where C is an arbitrary constant.

Indeed, $\Phi'(x) = [F(x) + C]' = F'(x) = f(x)$ for all $x \in (a, b)$. Therefore if the function $f(x)$ has an antiderivative on (a, b) then $f(x)$ has infinitely many antiderivatives on this interval.

The following theorem establishes the relationship between any two distinct antiderivatives of a given function.

Theorem 9.1. *If $F(x)$ and $\Phi(x)$ are any arbitrary antiderivatives of a given function $f(x)$ on the interval (a, b) then the difference of $F(x)$ and $\Phi(x)$ is equal to some constant so that $\Phi(x) - F(x) = C$, $x \in (a, b)$ (C is a constant).*

◄ Since $F(x)$ and $\Phi(x)$ are antiderivatives of $f(x)$ on (a, b) then $F'(x) = f(x)$ and $\Phi'(x) = f(x)$ for all $x \in (a, b)$. On differentiating $\varphi(x) = \Phi(x) - F(x)$ we obtain $\varphi'(x) = \Phi'(x) - F'(x) = f(x) - f(x) = 0$ for all $x \in (a, b)$. Let x_0 and x be two distinct points in (a, b). Then applying the mean value theorem (Theorem 8) to the function $\varphi(x)$ on the closed interval $[x_0, x]$, we get

$$\varphi(x) - \varphi(x_0) = (x - x_0)\varphi'(\xi),$$

where $x_0 < \xi < x$.

Since $\varphi'(x) = 0$ on (a, b) then $\varphi'(\xi) = 0$ and, consequently, $\varphi(x) = \varphi(x_0)$ for all $x \in (a, b)$. Hence, the function $\varphi(x)$ is constant on (a, b) and $\Phi(x) - F(x) = C$, where C is a constant for all $x \in (a, b)$. ►

Corollary. If $F(x)$ is any antiderivative of the function $f(x)$ on the interval (a, b) then any other antiderivative $\Phi(x)$ of $f(x)$ takes the form $\Phi(x) = F(x) + C$ where C is a constant.

The indefinite integral. The set of all antiderivatives of a function $f(x)$ on (a, b) is called the *indefinite integral* of $f(x)$ on (a, b), written $\int f(x)\,dx$. The symbol $\int$ is called the *integral sign*, the expression $f(x)\,dx$ is called the *element of integration*, the function $f(x)$ is called the *integrand* and the variable x is called the *variable of integration*.

If $F(x)$ is any antiderivative of the function $f(x)$ on the interval (a, b) then by virtue of the above corollary we have

$$\int f(x)\,dx = F(x) + C,$$

where C is an arbitrary constant.

Observe that any identity involving indefinite integrals in the left-hand and right-hand sides establishes the equality between sets of antiderivatives. This equality means that the sets contain the same elements, i.e., the same antiderivatives.

Sometimes the expression $\int f(x)\,dx$ is thought of as any element of the set of antiderivatives, i.e., any antiderivatives of the function $f(x)$.

We shall prove elsewhere the following theorem on the existence of the indefinite integral of a given function.

Theorem 9.2. *Let $f(x)$ be continuous on the interval (a, b). Then the function $f(x)$ possesses an antiderivative and, consequently, an indefinite integral on (a, b).*

The operation of computing the antiderivative or the indefinite integral of a function $f(x)$ is called the *integration* of $f(x)$. Clearly, this operation is the inverse of differentiation.

Properties of the indefinite integral. In what follows we assume that all functions involved are continuous on the interval (a, b); hence, we assume that the functions possess indefinite integrals on (a, b).

(1) The element of integration is equal to the differential of the indefinite integral so that

$$d\left(\int f(x)\,dx\right) = f(x)\,dx.$$

◄ Indeed, since $F'(x) = f(x)$ for all $x \in (a, b)$ then

$$d\left(\int f(x)\,dx\right) = d[F(x) + C] = dF(x) = F'(x)\,dx = f(x)\,dx. \blacktriangleright$$

(2) The integrand is equal to the derivative of the indefinite integral:

$$\left(\int f(x)\,dx\right)' = f(x).$$

◄ This readily follows from property (1). ►

(3) The indefinite integral of the differential of some function differs from this function by an arbitrary constant:

$$\int dF(x) = F(x) + C.$$

◄ Indeed, if $F'(x) = f(x)$ for all $x \in (a, b)$ then

$$\int dF(x) = \int F'(x)\,dx = \int f(x)\,dx = F(x) + C. \blacktriangleright$$

(4) If the integrand is multiplied by a constant factor then the indefinite integral is multiplied by this factor:

$$\int Af(x)\,dx = A\int f(x)\,dx \quad (A = \text{const},\ A \neq 0).$$

◄ Indeed, by virtue of property (2) we have

$$\left(\int Af(x)\,dx\right)' = Af(x) \text{ and } \left(A\int f(x)\,dx\right)' = A\left(\int f(x)\,dx\right)' = Af(x)$$

so that $\int Af(x)\,dx$ and $A\int f(x)\,dx$ correspond to the same set of antiderivatives of the function $Af(x)$. ►

(5) The indefinite integral of a sum (difference) of two functions is equal to the sum (difference) of the indefinite integrals of these functions:

$$\int [f(x) \pm \varphi(x)]\,dx = \int f(x)\,dx \pm \int \varphi(x)\,dx.$$

◄ By virtue of property (2) we have

$$\left(\int [f(x) \pm \varphi(x)]\,dx\right)' = f(x) \pm \varphi(x).$$

On the other hand we have

$$\left(\int f(x)\,dx \pm \int \varphi(x)\,dx\right)' = \left(\int f(x)\,dx\right)' \pm \left(\int \varphi(x)\,dx\right)' = f(x) \pm \varphi(x).$$

Therefore both $\int [f(x) \pm \varphi(x)]\,dx$ and $\int f(x)\,dx \pm \int \varphi(x)\,dx$ are antiderivatives of the function $f(x) \pm \varphi(x)$. Hence, they differ from each other by some constant C. ►

Corollary.

$$\int\left[\sum_{k=1}^{n} A_k f_k(x)\right] dx = \sum_{k=1}^{n} A_k \int f_k(x)\,dx,$$

where $A_k = \text{const}$ $(k = 1, 2, \ldots, n)$.

The expression of the form

$$\sum_{k=1}^{n} A_k f_k(x) = A_1 f_1(x) + A_2 f_2(x) + \ldots + A_n f_n(x),$$

where all A_k are some constants, is called the *linear combination of the functions* $f_k(x)$, $k = 1, 2, \ldots, n$. Therefore we can formulate this corollary in the following way: the indefinite integral of a linear combination of any finite number of functions is equal to the linear combination of the indefinite integrals of these functions.

Properties (4) and (5) are sometimes called the *linear properties of an indefinite integral.*

Some indefinite integrals involving elementary functions. Every formula for a derivative of a given function, i.e., a formula of the form $F'(x) = f(x)$ can be represented in terms of the function $f(x)$ and its antiderivative $F(x)$ as $\int f(x)\,dx = F(x) + C$ where C is an arbitrary constant. This enables us to derive the integration formulas directly implied by the differentiation formulas for basic elementary functions. These are:

$$\int x^{\alpha} dx = \frac{x^{\alpha+1}}{\alpha+1} + C,\ \alpha \neq -1,\ x > 0.$$

$$\int \frac{dx}{x} = \ln|x| + C,\ x \neq 0.$$

$$\int a^x dx = \frac{a^x}{\ln a} + C,\ 0 < a \neq 1.$$

(In particular, putting $a = \mathrm{e}$ gives $\int \mathrm{e}^x dx = \mathrm{e}^x + C$.)

$$\int \sin x\,dx = -\cos x + C.$$

$$\int \cos x\,dx = \sin x + C.$$

$$\int \frac{dx}{\sin^2 x} = -\cot x + C,\ x \neq n\pi\ (n = 0, \pm 1, \pm 2, \ldots).$$

$$\int \frac{dx}{\cos^2 x} = \tan x + C,\ x \neq \frac{\pi}{2} + n\pi\ (n = 0, \pm 1, \pm 2, \ldots).$$

$$\int \frac{dx}{\sqrt{1-x^2}} = \sin^{-1} x + C, \quad -1 < |x| < 1.$$

$$\int \frac{dx}{\sqrt{a^2-x^2}} = \sin^{-1} \frac{x}{a} + C, \quad |x| < a.$$

$$\int \frac{dx}{\sqrt{x^2 \pm a^2}} = \ln \left| x + \sqrt{x^2 \pm a^2} \right| + C$$ (the "minus" sign requires the condition $|x| > |a|$ be fulfilled).

$$\int \frac{dx}{1+x^2} = \tan^{-1} x + C.$$

$$\int \frac{dx}{a^2+x^2} = \frac{1}{a} \tan^{-1} \frac{x}{a} + C, \quad a \neq 0.$$

$$\int \sinh x \, dx = \cosh x + C.$$

$$\int \cosh x \, dx = \sinh x + C.$$

$$\int \frac{dx}{\cosh^2 x} = \tanh x + C.$$

$$\int \frac{dx}{\sinh^2 x} = -\coth x + C, \quad x \neq 0.$$

These formulas are easily verified by differentiating their right-hand sides whose derivatives are equal to the respective integrands.

The distinction between the operations of differentiation and integration is that on differentiating elementary functions we always obtain elementary functions while subjecting some elementary functions to integration can sometimes lead to nonelementary functions, i.e., there exist elementary functions which are nonintegrable in elementary functions. For example, it has been established that the following integrals cannot be represented in terms of elementary functions although, due to continuity of integrands, these indefinite integrals exist in the domains of definition:

$$\int e^{-x^2} dx, \qquad \int \frac{\sin x}{x} dx, \quad x \neq 0,$$

$$\int \sin x^2 dx, \qquad \int \frac{\cos x}{x} dx, \quad x \neq 0,$$

$$\int \cos x^2 dx, \qquad \int \frac{dx}{\ln x}, \quad 0 < x \neq 1.$$

Some indefinite integrals can be reduced, by suitable manipulations on their integrands, to integrals easily computed by applying the integration formulas that we have derived for basic elementary functions. To illustrate this method of integration we turn our attention to the examples of computing each of the following indefinite integrals.

Examples. (1) $\int\left(\sqrt{x^3}-\frac{1}{\sqrt{x^3}}\right)^2 dx.$

◀ Manipulating on the integrand we get

$$\int\left(\sqrt{x^3}-\frac{1}{\sqrt{x^3}}\right)^2 dx=\int\left(x^3-2+\frac{1}{x^3}\right)dx=\int(x^3-2+x^{-3})\,dx$$

$$=\int x^3dx-2\int dx+\int x^{-3}dx=\frac{x^4}{4}-2x+\frac{x^{-2}}{-2}+C=\frac{x^4}{4}-2x-\frac{1}{2x^2}+C.$$ ▶

(2) $\int\frac{(1+x^2)}{x^3+x}\,dx.$

◀ We have

$$\int\frac{(1+x)^2}{x^3+x}\,dx=\int\frac{1+2x+x^2}{x(1+x^2)}\,dx=\int\frac{(1+x^2)+2x}{x(1+x^2)}\,dx$$

$$=\int\left(\frac{1}{x}+\frac{2}{1+x^2}\right)dx=\int\frac{dx}{x}+2\int\frac{dx}{1+x^2}=\ln|x|+2\tan^{-1}x+C.$$ ▶

(3) $\int\frac{2\times 3^x+3\times 2^x}{5^x}\,dx.$

◀ We have

$$\int\frac{2\times 3^x+3\times 2^x}{5^x}\,dx=2\int\left(\frac{3}{5}\right)^x dx+3\int\left(\frac{2}{5}\right)^x dx$$

$$=2\,\frac{(3/5)^x}{\ln 3/5}+3\,\frac{(2/5)^x}{\ln 2/5}+C.$$ ▶

9.2 Methods of Integration

Integration by substitution. We wish to compute the indefinite integral $\int f(x)\,dx$ of the continuous function $f(x)$. Suppose that there exists a function $x=\varphi(t)$ having a continuous derivative $\varphi'(t)$ and the inverse function $t=\psi(x)$. Then we may write

$$\int f(x)\,dx=\int f[\varphi(t)]\,\varphi'(t)\,dt.$$

Clearly, substituting $t=\psi(x)$ into the right-hand side of the above identity we shall get the integral in terms of the original variable again.

◄ To prove the above identity we differentiate the left- and right-hand integrals with respect to x. On differentiating the left-hand integral, we get

$$\left(\int f(x)\,dx\right)'_x = f(x).$$

Observe that $t = \psi(x)$ is the inverse function of x so that

$$t'_x = \frac{1}{x'_t} = \frac{1}{\varphi'(t)},$$

where $\varphi'(t) \neq 0$.

Then applying to the right-hand integral the chain rule of differentiating a composite function, we obtain

$$\left(\int f[\varphi(t)]\,\varphi'(t)\,dt\right)'_x = \left(\int f[\varphi(t)]\varphi'(t)\,dt\right)'_t \cdot t'_x$$
$$= f[\varphi(t)]\varphi'(t)\,\frac{1}{\varphi'(t)} = f[\varphi(t)] = f(x).$$

Since the left- and right-hand integrals have the same derivative these integrals define the same set of antiderivatives of the function $f(x)$. Whence it follows that the desired identity is true. ►

The indentity being considered lays down the method of integration called *integration by substitution*. When using this method we have to choose a suitable function $\varphi(t)$ to simplify the computation of the original integral.

Examples. Compute the following integrals.

(1) $\displaystyle\int \frac{dx}{\sqrt{x^2 + a^2}}$ $(a > 0)$.

◄ Putting $x = a \sinh t$, we get $dx = a \cosh t\,dt$. Then

$$\int \frac{dx}{\sqrt{x^2+a^2}} = \int \frac{a\cosh t\,dt}{\sqrt{a^2(\sinh^2 t+1)}} = \int \frac{a\cosh t\,dt}{a\cosh t} = \int dt = t + \tilde{C}.$$

To express the result in terms of x we solve the equation $x = a\sinh t$ to get t as a function of x. Since $\sinh t = \dfrac{e^t - e^{-t}}{2}$, $a\dfrac{e^t - e^{-t}}{2} = x$ or $ae^{2t} - 2xe^t - a = 0$. Whence we get

$$e^t = \frac{x \pm \sqrt{x^2 + a^2}}{a}.$$

Since $e^t > 0$ we have $e^t = \dfrac{x + \sqrt{x^2+a^2}}{a}$ and $t = \ln(x + \sqrt{x^2+a^2}) - \ln a$. Finally, we obtain

$$\int \frac{dx}{\sqrt{x^2 + a^2}} = \ln (x + \sqrt{x^2 + a^2}) + C,$$

where $C = \tilde{C} - \ln a$. ►

(2) $\displaystyle\int \frac{dx}{(x + 2)\sqrt{x + 1}}$.

◄ Put $x = t^2 - 1$ so that $dx = 2t\,dt$ and $t = \sqrt{x + 1}$. Then

$$\int \frac{dx}{(x + 2)\sqrt{x + 1}} = \int \frac{2t\,dt}{(t^2 + 1)t} = 2\int \frac{dt}{t^2 + 1} = 2\tan^{-1}t + C.$$

Substituting $t = \sqrt{x + 1}$ into the right-hand side, we get

$$\int \frac{dx}{(x + 2)\sqrt{x + 1}} = 2\tan^{-1}\sqrt{x + 1} + C. \ ►$$

Remark. Suppose that in the integral $\int f(x)\,dx$ the element of integration $f(x)\,dx$ admits a representation of the form

$$f(x)\,dx = g\,[\psi(x)]\,\psi'(x)\,dx$$

so that

$$f(x)\,dx = g\,[\psi(x)]\,d[\psi(x)].$$

Assume that the function $g(t)$ is easily integrable, i.e., the integral

$$\int g(t)\,dt = F(t) + C$$

is easily computed. Then substituting $t = \psi(x)$ into the latter integral, we obtain

$$\int f(x)\,dx = F\,[\psi(x)] + C.$$

(3) $\displaystyle\int \frac{2^x - 2^{-x}}{2^x + 2^{-x}}\,dx.$

◄ Put $t = 2^x + 2^{-x}$ $(t > 0)$ so that $dt = (2^x \ln 2 - 2^{-x}\ln 2)\,dx$ and $(2^x - 2^{-x})\,dx = \dfrac{dt}{\ln 2}$. Then

$$\int \frac{2^x - 2^{-x}}{2^x + 2^{-x}}\,dx = \frac{1}{\ln 2}\int \frac{dt}{t} = \frac{1}{\ln 2}\ln t + C = \frac{\ln (2^x + 2^{-x})}{\ln 2} + C. \ ►$$

(4) $\displaystyle\int \frac{e^{2x}dx}{\sqrt{e^x + 1}}$.

◄ Putting $\sqrt{e^x + 1} = t$, we obtain $\dfrac{e^x dx}{\sqrt{e^x + 1}} = 2\,dt$ and $e^x = t^2 - 1$.

Then

$$\int \frac{e^{2x}dx}{\sqrt{e^x+1}} = \int e^x \frac{e^x dx}{\sqrt{e^x+1}} = 2\int (t^2-1)\,dt$$

$$= 2\left(\frac{t^3}{3} - t\right) + C = \frac{2}{3}(e^x+1)^{3/2} - 2\sqrt{e^x+1} + C$$

$$= \frac{2}{3}(e^x-2)\sqrt{e^x+1} + C. \ ►$$

Integration by parts. Let the functions $u = u(x)$ and $v = v(x)$ have continuous derivatives $u'(x)$ and $v'(x)$. Then applying the product rule of differentiation, we have

$$[u(x)\,v(x)]' = u(x)\,v'(x) + v(x)u'(x)$$

so that the product $u(x)\,v(x)$ of the given functions is an antiderivative of the sum $u(x)\,v'(x) + v(x)\,u'(x)$. Hence

$$\int [u(x)\,v'(x) + v(x)u'(x)]\,dx = u(x)\,v(x) + C.$$

By virtue of the linear property of indefinite integrals this expression can be written as

$$\int u(x)\,v'(x)\,dx = u(x)\,v(x) - \int v(x)\,u'(x)\,dx + C$$

or

$$\int u\,dv = uv - \int v\,du + C$$

since $v'(x)\,dx = dv$ and $u'(x)\,dx = du$ by the definition of the differential. Noting that the constant C can be combined with either of the integrals involved we may write the above identity as

$$\int u\,dv = uv - \int v\,du.$$

This identity lays down the method of integration called *integration by parts.* The idea is to replace the original integral $\int u\,dv$ by the integral $\int v\,du$ easily computed in some specific instances. The method involves splitting the original element of integration into two factors u and $dv = v'\,dx$ and subsequent differentiation of the former and integration of the latter. Clearly, integration by parts is helpful and constructive when this splitting provides easily differentiable u and easily integrable dv.

To illustrate integration by parts we shall work through the integral $\int (2-3x)\cos x\,dx$.

◄ Here $u\,dv = (2-3x)\cos x\,dx$. Put $u = 2-3x$ and $dv = \cos x\,dx$ so that $du = -3dx$ and $v = \int \cos x\,dx = \sin x$. Then

$$\int u\,dv = uv - \int v\,du = \int (2-3x)\cos x\,dx$$

$$= (2-3x)\sin x + 3\int \sin x\,dx = (2-3x)\sin x - 3\cos x + C. \ ►$$

Notice that if we put $u = \cos x$ and $dv = (2 - 3x)\,dx$ or $u = (2 - 3x)\cos x$ and $dv = dx$ the integral $\int v\,du$ becomes more complicated than the original one.

On integrating dv to get the function v we may choose any suitable value of the constant of integration $\tilde{C}$ since $\tilde{C}$ is not involved in the final result (to verify this it suffices to substitute $v + \tilde{C}$ for v in the identity describing integration by parts). In particular, it is convenient to put $\tilde{C} = 0$.

Examples. **(1)** We wish to compute the integral $\int \ln x\,dx$.

◄ Noting that $u\,dv = \ln x\,dx$ we have the unique factorization $u = \ln x$ and $dv = dx$. Then $du = \frac{dx}{x}$ and $v = \int dx = x$; whence we obtain

$$\int \ln x\,dx = x\ln x - \int dx = x\ln x - x + C = x(\ln x - 1) + C. \blacktriangleright$$

(2) Integrate by parts $\int \sqrt{a^2 - x^2}\,dx$ $(|x| < |a|)$.

◄ Put $u = \sqrt{a^2 - x^2}$ and $dv = dx$ so that $du = -\frac{x\,dx}{\sqrt{a^2 - x^2}}$ and $v = x$. Integrating by parts, we get

$$\int \sqrt{a^2 - x^2}\,dx = x\sqrt{a^2 - x^2} + \int \frac{x^2\,dx}{\sqrt{a^2 - x^2}} + 2C.$$

On manipulating with the integrand of the right-hand integral, we obtain

$$\begin{aligned}\int \sqrt{a^2 - x^2}\,dx &= x\sqrt{a^2 - x^2} + \int \frac{a^2 - (a^2 - x^2)}{\sqrt{a^2 - x^2}}\,dx + 2C\\ &= x\sqrt{a^2 - x^2} + a^2\int \frac{dx}{\sqrt{a^2 - x^2}} - \int \sqrt{a^2 - x^2}\,dx + 2C\\ &= x\sqrt{a^2 - x^2} + a^2\sin^{-1}\frac{x}{a} - \int \sqrt{a^2 - x^2}\,dx + 2C.\end{aligned}$$

We have obtained the equation in one unknown $\int \sqrt{a^2 - x^2}\,dx$. Solving this equation, we get

$$\int \sqrt{a^2 - x^2}\,dx = \frac{1}{2}x\sqrt{a^2 - x^2} + \frac{a^2}{2}\sin^{-1}\frac{x}{a} + C. \blacktriangleright$$

Problem. Verify the following formulas

(a) $\int \sqrt{x^2 + a^2}\,dx = \frac{1}{2}x\sqrt{x^2 + a^2} + \frac{a^2}{2}\ln(x + \sqrt{x^2 + a^2}) + C;$

(b) $\int \sqrt{x^2 - a^2}\,dx = \frac{1}{2}x\sqrt{x^2 - a^2} - \frac{a^2}{2}\ln|x + \sqrt{x^2 - a^2}| + C,$

where $|x| > |a|$.

Remark. Sometimes we can use integration by parts to compute the right-hand integral $\int v\,du$. To illustrate this approach we shall work through the following examples of integrating by parts.

Examples. (1) We wish to compute the integral $\int x^2 2^x\,dx$.

◀ Put $u = x^2$ and $dv = 2^x\,dx$ so that $du = 2x\,dx$ and $v = \frac{2^x}{\ln 2}$.
Integration by parts gives

$$\int x^2 2^x\,dx = x^2\,\frac{2^x}{\ln 2} - \frac{2}{\ln 2}\int x\,2^x\,dx.$$

The right-hand integral can also be integrated by parts. If we put $u = x$ and $dv = 2^x\,dx$ then $du = dx$ and $v = \frac{2^x}{\ln 2}$ so that

$$\begin{aligned}\int x^2 2^x\,dx &= x^2\,\frac{2^x}{\ln 2} - \frac{2}{\ln 2}\left(x\,\frac{2^x}{\ln 2} - \frac{1}{\ln 2}\int 2^x\,dx\right)\\ &= x^2\,\frac{2^x}{\ln 2} - \frac{2}{\ln 2}\left(x\,\frac{2^x}{\ln 2} - \frac{2^x}{\ln^2 2}\right) + C\\ &= \left(x^2 - \frac{2}{\ln 2}x + \frac{2}{\ln^2 2}\right)\frac{2^x}{\ln 2} + C. \text{ ▶}\end{aligned}$$

(2) Compute the integral $\int e^{\alpha x}\cos\beta x$ $(\alpha \neq 0,\ \beta \neq 0)$.

◀ Applying integration by parts, we put $u = e^{\alpha x}$ and $dv = \cos\beta x\,dx$ (or $u = \cos\beta x$ and $dv = e^{\alpha x}\,dx$) so that

$$du = \alpha e^{\alpha x}\,dx \text{ and } v = \int\cos\beta x\,dx = \frac{\sin\beta x}{\beta}.$$

Then

$$\int e^{\alpha x}\cos\beta x\,dx = \frac{1}{\beta}\,e^{\alpha x}\sin\beta x - \frac{\alpha}{\beta}\int e^{\alpha x}\sin\beta x\,dx.$$

Integrating the latter integral by parts gives: $u = e^{\alpha x}$ and $dv = \sin\beta x\,dx$ so that $du = \alpha e^{\alpha x}dx$ and $v = -\frac{\cos\beta x}{\beta}$ and

$$\int e^{\alpha x}\sin\beta x\,dx = -\frac{1}{\beta}\,e^{\alpha x}\cos\beta x + \frac{\alpha}{\beta}\int e^{\alpha x}\cos\beta x\,dx.$$

Substituting this identity into the preceding one, we obtain

$$\int e^{\alpha x}\cos\beta x\,dx = \frac{1}{\beta}\,e^{\alpha x}\sin\beta x + \frac{\alpha}{\beta^2}\,e^{\alpha x}\cos\beta x - \frac{\alpha^2}{\beta^2}\int e^{\alpha x}\cos\beta x\,dx.$$

Thus applying integration by parts twice we have arrived at the equation in one unknown, whence we have

$$\left(1 + \frac{\alpha^2}{\beta^2}\right)\int e^{\alpha x}\cos\beta x\,dx = \frac{e^{\alpha x}}{\beta^2}\,(\alpha\cos\beta x + \beta\sin\beta x)$$

and

$$\int e^{\alpha x}\cos\beta x\,dx = \frac{e^{\alpha x}}{\alpha^2+\beta^2}(\alpha\cos\beta x + \beta\sin\beta x) + C.$$

Analogously, we can compute the integral

$$\int e^{\alpha x}\sin\beta x\,dx = \frac{e^{\alpha x}}{\alpha^2+\beta^2}(\alpha\sin\beta x - \beta\cos\beta x) + C. \blacktriangleright$$

Integration by parts is helpful to compute a number of integrals given below.

Integrals of the form $\int P_n(x)\ln x\,dx$, where $P_n(x) = \sum_{k=0}^{n} c_k x^k$ is the nth degree polynomial. If we put $u = \ln x$ and $dv = P_n(x)\,dx$ then $du = \frac{dx}{x}$ and $v = \int P_n(x)\,dx = Q_{n+1}(x)$, where $Q_{n+1}(x)$ is a polynomial of degree $(n+1)$, and integration by parts gives

$$\begin{aligned}\int P_n(x)\ln x\,dx &= Q_{n+1}(x)\ln x - \int \frac{Q_{n+1}(x)}{x}\,dx\\ &= Q_{n+1}(x)\ln x - H_{n+1}(x) + C,\end{aligned}$$

where $H_{n+1}(x)$ is a polynomial of degree $(n+1)$.

Example. Compute the integral $\int(4x^3+2x)\ln x\,dx$.

◄ Put $u = \ln x$ and $dv = (4x^3+2x)\,dx$ so that $du = \frac{dx}{x}$ and $v = \int(4x^3 + 2x)\,dx = x^4 + x^2$. Then integration by parts gives

$$\begin{aligned}\int(4x^3+2x)\ln x\,dx &= (x^4+x^2)\ln x - \int(x^3+x)\,dx\\ &= (x^4+x^2)\ln x - \frac{x^4}{4} - \frac{x^2}{2} + C. \blacktriangleright\end{aligned}$$

Integrals of the form $\int P_n(x)\tan^{-1}\alpha x\,dx$ *and* $\int P_n(x)\cot^{-1}\alpha x\,dx$, where α is a real number. These integrals are reduced to those of rational functions. For example, if we put $u = \tan^{-1}\alpha x$ and $dv = P_n(x)\,dx$ in the former integral then $du = \frac{\alpha\,dx}{1+\alpha^2x^2}$ and $v = Q_{n+1}(x)$ and integration by parts gives

$$\int P_n(x)\tan^{-1}\alpha x\,dx = Q_{n+1}(x)\tan^{-1}\alpha x - \alpha\int\frac{Q_{n+1}(x)}{1+\alpha^2x^2}\,dx.$$

Analogously, we work through the latter integral.

Example. Compute the integral $\int(3x^2+1)\tan^{-1}x\,dx$.

◄ Put $u = \tan^{-1}x$ and $dv = (3x^2+1)\,dx$ so that $du = \frac{dx}{1+x^2}$ and

$v = x^3 + x$. Then integration by parts gives

$$\int (3x^2 + 1)\tan^{-1} x\,dx = (x^3 + x)\tan^{-1} x - \int \frac{x^3 + x}{1 + x^2}\,dx$$
$$= (x^3 + x)\tan^{-1} x - \int x\,dx = (x^3 + x)\tan^{-1} x - \frac{x^2}{2} + C. \blacktriangleright$$

Integrals of the form $\int P_n(x)\sin^{-1}\alpha x\,dx$ *and* $\int P_n(x)\cos^{-1}\alpha x\,dx$, where α is a real number. To integrate these integrals by parts we put $u = \sin^{-1}\alpha x$ ($u = \cos^{-1}\alpha x$ in the latter integral) and $dv = P_n(x)\,dx$. Then

$$du = \frac{\alpha\,dx}{\sqrt{1 + \alpha^2 x^2}} \quad \left(du = -\frac{\alpha\,dx}{\sqrt{1 - \alpha^2 x^2}}\right), \quad v = Q_{n+1}(x)$$

and

$$\int P_n(x)\sin^{-1}\alpha x\,dx = Q_{n+1}(x)\sin^{-1}\alpha x - \alpha\int \frac{Q_{n+1}(x)}{\sqrt{1 - \alpha^2 x^2}}\,dx.$$

The integral in the right-hand side can be computed by using different methods which will be discussed in detail in Sec. 9.4.

Example. Integrate by parts $\int x^2 \sin^{-1} x\,dx$.

◄ Put $u = \sin^{-1} x$ and $dv = x^2\,dx$. Then $du = \frac{dx}{\sqrt{1 - x^2}}$ and $v = \frac{x^3}{3}$. Integrating by parts, we get

$$\int x^2 \sin^{-1} x\,dx = \frac{x^3}{3}\sin^{-1} x - \frac{1}{3}\int \frac{x^3\,dx}{\sqrt{1 - x^2}}.$$

By the substitution $t = \sqrt{1 - x^2}$ so that $x^2 = 1 - t^2$ and $x\,dx = -t\,dt$ the integral on the right is reduced to

$$\int \frac{x^3\,dx}{\sqrt{1 - x^2}} = \int \frac{x^2}{\sqrt{1 - x^2}}\,x\,dx = -\int \frac{1 - t^2}{t}\,t\,dt$$
$$= \int (t^2 - 1)dt = \frac{1}{3}t^3 - t + C = -\frac{1}{3}(x^2 + 2)\sqrt{1 - x^2} + C.$$

Finally we have

$$\int x^2 \sin^{-1} x\,dx = \frac{x^3}{3}\sin^{-1} x + \frac{1}{9}(x^2 + 2)\sqrt{1 - x^2}. \blacktriangleright$$

Integrals of the form $\int P_n(x)\,e^{\lambda x}\,dx$ where λ is a real number. Putting $u = P_n(x)$ and $dv = e^{\lambda x}\,dx$ we obtain $du = P_n'(x)\,dx$ and $v = \frac{1}{\lambda}e^{\lambda x}$ ($\lambda \neq 0$)

so that

$$\int P_n(x)\,e^{\lambda x}\,dx = \frac{1}{\lambda}\,P_n(x)\,e^{\lambda x} - \frac{1}{\lambda}\int P_n'(x)\,e^{\lambda x}\,dx,$$

where $P_n'(x)$ is a polynomial of degree $(n-1)$.

Then applying integration by parts to the right-hand integral we again obtain the expression which can also be subjected to integration by parts. Continuing this process n times we finally arrive at the integral

$$\int e^{\lambda x}\,dx = \frac{1}{\lambda}\,e^{\lambda x} + C.$$

Example. Compute the integral $\int (x^2 + 2x)\,e^x dx$.

◀ Put $u = x^2 + 2x$ and $dv = e^x dx$ so that $du = 2(x+1)\,dx$ and $v = e^x$. Then integration by parts gives

$$\int (x^2 + 2x)\,e^x dx = (x^2 + 2x)\,e^x - 2\int (x+1)\,e^x dx.$$

To integrate the right-hand integral by parts we put $u = x + 1$ and $dv = e^x dx$. Then $du = dx$ and $v = e^x$ so that

$$\int (x+1)\,e^x dx = (x+1)\,e^x - \int e^x dx = (x+1)\,e^x - e^x + C = xe^x + C.$$

Finally, we have

$$\int (x^2 + 2x)\,e^x dx = (x^2 + 2x)\,e^x - 2xe^x + C = x^2 e^x + C. \blacktriangleright$$

Integrals of this form can also be computed by the *method of comparing* (unknown) *coefficients*. In this case we assume that the integral is equal to the product of the nth degree polynomial

$$Q_n(x) = b_0 + b_1 x + \ldots + b_n x^n$$

with unknown coefficients $b_0, b_1, \ldots, b_n$ by the function $e^{\lambda x}$, i.e., we assume that

$$\int P_n(x)\,e^{\lambda x} dx = Q_n(x)\,e^{\lambda x}.$$

On differentiating the integral we get

$$P_n(x)\,e^{\lambda x} = Q_n'(x)\,e^{\lambda x} + \lambda Q_n(x)\,e^{\lambda x}$$

whence, dividing by $e^{\lambda x} \neq 0$, we get

$$P_n(x) = Q_n'(x) + \lambda\,Q_n(x).$$

Since the polynomial on the left is to be identically equal to the sum of polynomials on the right, the coefficients of any power of x must be equal; so comparing coefficients, we obtain the system of $(n+1)$ linear equations in n unknowns b_k. This system has a unique solution, for the respective determinant is distinct from zero. To demonstrate how this method works we shall compute $\int (x^2 + 2x)\,e^x dx$.

◄ Put

$$\int (x^2 + 2x)\,e^x dx = (b_0 + b_1 x + b_2 x^2)\,e^x,$$

where b_0, b_1, b_2 are unknown coefficients.

On differentiating this identity, we get

$$(x^2 + 2x)\,e^x = (b_1 + 2b_2 x)\,e^x + (b_0 + b_1 x + b_2 x^2)\,e^x.$$

Whence, division by $e^x \neq 0$ gives

$$2x + x^2 = b_0 + b_1 x + b_2 x + b_1 + 2b_2 x$$

or

$$2x + x^2 = (b_0 + b_1) + (b_1 + 2b_2)\,x + b_2 x^2.$$

Comparing the coefficients of equal powers of x, we arrive at the system of linear equations

$$\begin{array}{c|l} x^0 & 0 = b_0 + b_1 \\ x^1 & 2 = b_1 + 2b_2 \\ x^2 & 1 = b_2 \end{array} \Bigg\},$$

whose solution is $b_0 = 0$, $b_1 = 0$ and $b_2 = 1$.

Hence the original integral becomes

$$\int (x^2 + 2x)\,e^x dx = x^2 e^x + C. \ ►$$

Integrals of the form $\int P_n(x) \sin \beta x\, dx$ and $\int P_n(x) \cos \beta x\, dx$, where β is a real number, $\beta \neq 0$.

◄ Putting $u = P_n(x)$ and $dv = \sin \beta x\, dx$ ($dv = \cos \beta x\, dx$ in the latter integral), we have $du = P_n'(x)\, dx$ and $v = -\dfrac{\cos \beta x}{\beta}$ $\left(v = \dfrac{\sin \beta x}{\beta}\right)$. Then

$$\int P_n(x) \sin \beta x\, dx = -P_n(x)\,\frac{\cos \beta x}{\beta} + \frac{1}{\beta}\int P_n'(x) \cos \beta x\, dx.$$

Applying integration by parts n times, we finally obtain

$$\int \sin \beta x\, dx = -\frac{\cos \beta x}{\beta} \quad \text{and} \quad \int \cos \beta x\, dx = \frac{\sin \beta x}{\beta}. \ ►$$

Example. Compute the integral $\int (x^2 - 1) \cos x\, dx$.

◄ Put $u = x^2 - 1$ and $dv = \cos x\, dx$ so that $du = 2x\, dx$ and $v = \sin x$. Then integration by parts gives

$$\int (x^2 - 1) \cos x\, dx = (x^2 - 1) \sin x - 2\int x \sin x\, dx.$$

Now we put $u = x$ and $dv = \sin x\, dx$ so that $du = dx$ and $v = -\cos x$. Then we have

$$\begin{aligned}\int (x^2 - 1) \cos x\, dx &= (x^2 - 1) \sin x + 2x \cos x - 2 \sin x + C \\ &= (x^2 - 3) \sin x + 2x \cos x + C. \ ►\end{aligned}$$

Integration by parts can also be helpful in dealing with some other integrals apart from those considered above. By way of illustration we shall compute the integral $\int \frac{x\,dx}{\sin^2 x}$.

◄ Putting $u = x$ and $dv = \frac{dx}{\sin^2 x}$, we obtain $du = dx$ and $v = \int \frac{dx}{\sin^2 x} = -\cot x$ so that

$$\int \frac{x\,dx}{\sin^2 x} = -x\cot x + \int \cot x\,dx.$$

By the substitution $t = \sin x$ and $dt = \cos x\,dx$ the integral on the right is reduced to

$$\int \cot x\,dx = \int \frac{\cos x\,dx}{\sin x} = \int \frac{dt}{t} = \ln|t| + C = \ln|\sin x| + C.$$

Finally, we have

$$\int \frac{x\,dx}{\sin^2 x} = -x\cot x + \ln|\sin x| + C. ►$$

9.3 Integrating Rational Functions

Rational functions. The simplest rational function is the nth polynomial

$$Q_n(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n,$$

where the coefficients $a_0, a_1, \ldots, a_n$ are real numbers and $a_0 \neq 0$.

The polynomial $Q_n(x)$ is said to be *monic* if $a_0 = 1$.

The real number b is called the *root* of the polynomial $Q_n(x)$ if $Q_n(b) = 0$.

It is known from algebra that any real polynomial $Q_n(x)$ can be uniquely factorized into a product of monic linear polynomial $x - b$ and monic quadratic polynomials $x^2 + px + q$, where p and q are real coefficients and every monic quadratic polynomial is irreducible to a product of linear polynomials, for it has no real roots. Then gathering equal factors if any, we can write the monic polynomial $Q_n(x)$ in the form

$$Q_n(x) = (x-a)^{\alpha}(x-b)^{\beta}\ldots(x-l)^{\lambda}(x^2+p_1x+q_1)^{\mu_1}\ldots(x^2+p_sx+q_s)^{\mu_s},$$

where the exponents $\alpha, \beta, \ldots, \lambda, \mu_1, \ldots, \mu_s$ are natural numbers.

It is easy to notice that for the nth degree monic polynomial $Q_n(x)$ the following condition is satisfied

$$\alpha + \beta + \ldots + \lambda + 2(\mu_1 + \mu_2 + \ldots + \mu_s) = n.$$

When $\alpha = 1$ the root a is said to be *unrepeated* or *simple*; when $\alpha \geqslant 2$ we say that a is a *repeated root* of Q_n and call the number α the *multiplicity* of a.

A *real rational function* $f(x)$ is the ratio $\dfrac{P_m(x)}{Q_n(x)}$ of two real polynomials $P_m(x)$ and $Q_n(x)$ which have no factor in common.

The rational function $f(x) = \dfrac{P_m(x)}{Q_n(x)}$ is called the *proper rational function* if the numerator $P_m(x)$ has a degree lower than that of the denominator $Q_n(x)$, i.e., if $m < n$. If $m \geqslant n$ we can apply the division algorithm to express $f(x)$ as

$$f(x) = \frac{P_m(x)}{Q_n(x)} = R_{m-n}(x) + \frac{\tilde{P}(x)}{Q_n(x)},$$

where $R_{m-n}(x)$ and $\tilde{P}(x)$ are real polynomials and $\dfrac{\tilde{P}(x)}{Q_n(x)}$ is a proper rational function. For example, if we apply the division algorithm to the rational function $\dfrac{x^5 + 1}{x^2 + 1} = \dfrac{P_5(x)}{Q_2(x)}$, we obtain

$$\begin{array}{rl}
 & x^3 - x \\
x^2 + 1 \big) & \overline{x^5 + 0x^4 + 0x^3 + 0x^2 + 0x + 1} \\
 & \underline{x^5 + 0x^4 + 1x^3} \\
 & \qquad\quad - 1x^3 + 0x^2 + 0x + 1 \\
 & \qquad\quad \underline{- 1x^3 - 0x^2 - 1x \qquad} \\
 & \qquad\qquad\qquad\qquad\qquad x + 1.
\end{array}$$

Hence, $\dfrac{x^5 + 1}{x^2 + 1} = x^3 - x + \dfrac{x + 1}{x^2 + 1}$ so that $R_3(x) = x^3 - x$, $\tilde{P}(x) = x + 1$ and $\dfrac{x + 1}{x^2 + 1}$ is a proper rational function.

Splitting a rational function into partial fractions. *Partial fractions* are proper rational functions of the form

$$\frac{A}{x - a}, \quad \frac{A}{(x - a)^k}, \quad \frac{Mx + N}{x^2 + px + q} \quad \text{and} \quad \frac{Mx + n}{(x^2 + px + q)^k},$$

where A, M, N, a, p, and q are real numbers, k ($k \geqslant 2$) is a natural number and quadratic trinomial $x^2 + px + q$ has no real root so that the condition $\dfrac{p^2}{4} - q < 0$ or $q - \dfrac{p^2}{4} > 0$ holds.

We turn once again to algebra and quote the following important fact.

Theorem. 9.3. *Let* $f(x) = \dfrac{P_m(x)}{Q_n(x)}$ $(m < n)$ *be a proper real rational function and let* $Q_n(x) = (x - a)^\alpha (x - b)^\beta \ldots (x^2 + p_s x + q_s)^{\mu_s}$. *Then* $f(x)$

is uniquely reducible to a sum of partial fractions of the form

$$\frac{P_m(x)}{Q_n(x)} = \frac{A_1}{x-a} + \frac{A_2}{(x-a)^2} + \dots + \frac{A_\alpha}{(x-a)^\alpha}$$
$$+ \frac{B_1}{x-b} + \frac{B_2}{(x-b)^2} + \dots + \frac{B}{(x-b)^\beta}$$
$$+\dots+ \frac{M_1x+N_1}{x^2+p_sx+q_s} + \frac{M_2x+N_2}{(x^2+p_sx+q_s)^2} +\dots+ \frac{M_{\mu_s}x+N_{\mu_s}}{(x^2+p_sx+q_s)^{\mu_s}},$$

where $A_1, A_2, \dots, A_\alpha, B_1, B_2, \dots, B_\beta, \dots, M_1, N_1, M_2, N_2, \dots, M_{\mu_s}, N_{\mu_s}$ *are real numbers not all equal to zero.*

To define the coefficients in the numerators of the partial fractions we multiply the left- and right-hand sides of the preceding expression by $Q_n(x)$ and apply the method of comparing (unknown) coefficients, i.e., we compare the coefficients of equal powers of x on the left and on the right thus obtaining a system of linear equations in the desired unknowns. The solution of this system uniquely defines the coefficients we are seeking for.

Sometimes it is convenient to use another method of computing coefficients of partial fractions. If two polynomials are identical the identity holds for any value of x; then multiplying both sides of the expansion given in Theorem 9.3 by $Q_n(x)$ and substituting some specific values for x into the identity thus obtained it is possible to get simple equations for the unknown coefficients. The method is often useful when $Q_n(x)$ has only simple real roots, substituting the values of the roots for x we arrive at simple equations for the unknown coefficients.

To illustrate these methods we use some specific examples.

Examples. (1) Split the proper rational function $\frac{3x^2-6x+2}{x^3-3x^2+2x}$ into partial fractions.

◀ On factoring the denominator, we get

$$(x^3-3x^2+2x) = x(x^2-3x+2) = x(x-1)(x-2).$$

Clearly, the denominator has real simple roots. Then we can write

$$\frac{3x^2-6x+2}{x^3-3x^2+2x} = \frac{A}{x} + \frac{B}{x-1} + \frac{C}{x-2}.$$

Multiplying both sides by the denominator, we obtain the identity

$$3x^2-6x+2 = A(x-1)(x-2) + Bx(x-2) + Cx(x-1) \qquad (*)$$

or

$$3x^2-6x+2 = (A+B+C)x^2 + (-3A-2B-C)x + 2A:$$

Applying the method of comparing coefficients, we arrive at the system of linear equations in the unknowns A, B, C

$$\begin{array}{l|l} x^2 & A + B + C = 3 \\ x^1 & -3A - 2B - C = -6 \\ x^0 & 2A = 2 \end{array} \left.\vphantom{\begin{array}{l}1\\1\\1\end{array}}\right\},$$

whose solution yields $A = 1$, $B = 1$ and $C = 1$.

We obtain the same result by substituting the values of the roots of the denominator for x in (*). Indeed, the roots are $x_1 = 0$, $x_2 = 1$ and $x_3 = 2$. Then the identity yields

$$\begin{aligned} 2 &= 2A \quad \text{and } A = 1 \text{ for } x = 0, \\ -1 &= -B \quad \text{and } B = 1 \text{ for } x = 1, \\ 2 &= 2C \quad \text{and } C = 1 \text{ for } x = 2. \end{aligned}$$

Hence

$$\frac{3x^2 - 6x + 2}{x^3 - 3x^2 + 2x} = \frac{1}{x} + \frac{1}{x - 1} + \frac{1}{x - 2} \blacktriangleright.$$

(2) Split the proper rational function $\dfrac{x^3 + 3x + 1}{x^5 + 3x^4 + 3x^3 + x^2}$ into partial fractions.

◀ The denominator can be factorized as

$$x^5 + 3x^4 + 3x^3 + x^2 = x^2(x^3 + 3x^2 + 3x + 1) = (x + 1)^3 x^2.$$

Thus it has the repeated root $x_1 = 0$ of multiplicity 2 and the repeated root $x_2 = -1$ of multiplicity 3. Then

$$\frac{x^3 + 3x + 1}{x^5 + 3x^4 + 3x^3 + x^2} = \frac{A_1}{x} + \frac{A_2}{x^2} + \frac{B_1}{x + 1} + \frac{B_2}{(x + 1)^2} + \frac{B_3}{(x + 1)^3}.$$

Multiplying out, as before, we get

$$\begin{aligned} x^3 + 3x + 1 &= A_1 x(x + 1)^3 + A_2(x + 1)^3 + B_1 x^2(x + 1)^2 \\ &\quad + B_2 x^2(x + 1) + B_3 x^2 \qquad (**) \end{aligned}$$

or

$$\begin{aligned} x^3 + 3x + 1 &= (A_1 + B_1)x^4 + (3A_1 + A_2 + 2B_1 + B_2)x^3 \\ &\quad + (3A_1 + 3A_2 + B_1 + B_2 + B_3)x^2 + (A_1 + 3A_2)x + A_2. \end{aligned}$$

The method of comparing coefficients gives

$$\begin{array}{l|l} x^4 & A_1 + B_1 = 0 \\ x^3 & 3A_1 + A_2 + 2B_1 + B_2 = 1 \\ x^2 & 3A_1 + 3A_2 + B_1 + B_2 + B_3 = 0 \\ x^1 & A_1 + 3A_2 = 3 \\ x^0 & A_2 = 1 \end{array} \left.\vphantom{\begin{array}{l}1\\1\\1\\1\\1\end{array}}\right\},$$

so that $A_1 = 0$, $A_2 = 1$, $B_1 = 0$, $B_2 = 0$, $B_3 = -3$ and

$$\frac{x^3 + 3x + 1}{x^5 + 3x^4 + 3x^3 + x^2} = \frac{1}{x^2} - \frac{3}{(x+1)^3}.$$

Proceeding as in the previous example, we put $x = 0$ and $x = -1$ in the identity (**). This yields $A_2 = 1$ for $x = 0$ and $B_3 = -3$ for $x = -1$. Substituting these values for A_2 and B_3 into (**), we get

$$x^3 + 3x + 1 = A_1x(x+1)^3 + (x+1)^3 + B_1x^2(x+1)^2 + B_2x^2(x+1) - 3x^2$$

or

$$x^3 + 3x + 1 - (x+1)^3 + 3x^2 = A_1x(x+1)^3 + B_1x^2(x+1)^2 + B_2x^2(x+1),$$

whence

$$0 = A_1x(x+1)^3 + B_1x^2(x+1)^2 + B_2x^2(x+1)$$

and

$$A_1(x+1)^2 + B_1x(x+1) + B_2x = 0.$$

Put $x = 0$ and $x = -1$. Then $A_1 = 0$, $B_2 = 0$ and, consequently, $B_1 = 0$. Thus we have the same values for the coefficients as those obtained by the method of comparing coefficients, namely,

$$A_1 = 0,\ A_2 = 1,\ B_1 = 0,\ B_2 = 0,\ B_3 = -3. \blacktriangleright$$

(3) Split the proper rational function $\frac{x^3 + x^2 + 1}{(x^2+1)^2}$ into partial fractions.

◄ The denominator has no real roots since $x^2 + 1$ does not vanish for any value of x. Then there must hold

$$\frac{x^3 + x^2 + 1}{(x^2+1)^2} = \frac{M_1x + N_1}{x^2 + 1} + \frac{M_2x + N_2}{(x^2+1)^2}.$$

Whence

$$x^3 + x^2 + 1 = (M_1x + N_1)(x^2 + 1) + M_2x + N_2$$

or

$$x^3 + x^2 + 1 = M_1x^3 + N_1x^2 + (M_1 + M_2)x + (N_1 + N_2).$$

Comparing the coefficients of equal powers of x, we get

$$M_1 = 1,\ N_1 = 1,\ M_1 + M_2 = 0,\ N_1 + N_2 = 1.$$

Whence

$$M_1 = 1,\ N_1 = 1,\ M_2 = -1,\ N_2 = 0$$

and, consequently,

$$\frac{x^3+x^2+1}{(x^2+1)^2}=\frac{x+1}{x^2+1}-\frac{x}{(x^2+1)^2}. \blacktriangleright$$

It is worth noting that sometimes it can be easier to arrive at the desired result without applying the method of comparing coefficients. For instance, in example (2) after a little algebra we obtain

$$\frac{x^3+3x+1}{x^5+3x^4+3x^3+x^2}=\frac{(x^3+3x^2+3x+1)-3x^2}{x^2(x+1)^3}$$
$$=\frac{(x+1)^3-3x^2}{x^2(x+1)^3}=\frac{1}{x^2}-\frac{3}{(x+1)^3}.$$

Integrating partial fractions. A rational function can be uniquely represented in the form of a sum of a polynomial (a zero polynomial in the case of a proper rational function) and a proper rational function which can be splitted into partial fractions. Since any real polynomial can be easily computed by applying standard integration formulas for elementary functions, integration of a rational function becomes integration of a sum of partial fractions. So now we turn our attention to techniques of integrating partial fractions.

Using standard integration formulas we can easily integrate partial fractions whose denominators are monic linear polynomials or power functions of them, namely

$$\int\frac{A\,dx}{x-a}=A\int\frac{d(x-a)}{x-a}=A\ln|x-a|+C.$$
$$\int\frac{A\,dx}{(x-a)^k}=A\int(x-a)^{-k}d(x-a)=\frac{A}{1-k}(x-a)^{-k+1}+C$$
$$=\frac{A}{(1-k)(x-a)^{k-1}}+C.$$

To integrate a partial fraction of the form $\dfrac{Mx+N}{x^2+px+q}$ we apply the method of *completing the square* in the denominator. This yields

$$x^2+px+q=\left[x^2+2x\frac{p}{2}+\left(\frac{p}{2}\right)^2\right]+q-\left(\frac{p}{2}\right)^2$$
$$=\left(x+\frac{p}{2}\right)^2+\left(q-\frac{p^2}{4}\right).$$

Since the second summand is positive we set it equal to a^2 where $a=+\sqrt{q-\dfrac{p^2}{4}}$. Then the substitution $x+\dfrac{p}{2}=t$ gives $dx=dt$ and

$x^2 + px + q = t^2 + a^2$. By virtue of the linear property of the indefinite integral we have

$$\begin{aligned}\int \frac{Mx+N}{x^2+px+q}\,dx &= \int \frac{M\left(t-\frac{p}{2}\right)+N}{t^2+a^2}\,dt\\ &= \frac{M}{2}\int\frac{2t\,dt}{t^2+a^2} + \left(N-\frac{Mp}{2}\right)\int\frac{dt}{t^2+a^2}\\ &= \frac{M}{2}\int\frac{d(t^2+a^2)}{t^2+a^2} + \left(N-\frac{Mp}{2}\right)\int\frac{dt}{t^2+a^2}\\ &= \frac{M}{2}\ln(t^2+a^2) + \left(N-\frac{Mp}{2}\right)\frac{1}{a}\tan^{-1}\frac{t}{a} + C\\ &= \frac{M}{2}\ln(x^2+px+q) + \frac{2N-Mp}{\sqrt{4q-p^2}}\tan^{-1}\frac{2x+p}{\sqrt{4q-p^2}} + C.\end{aligned}$$

To illustrate the method of integrating partial fractions of this form we work through $\int \frac{2-x}{x^2+4x+6}\,dx$.

◄ Clearly, the quadratic polynomial $x^2 + 4x + 6$ has no real roots since $\frac{p^2}{4} - q = -2 < 0$. Completing the square, we get

$$x^2 + 4x + 6 = (x^2 + 4x + 4) + 2 = (x + 2)^2 + 2.$$

The substitution $x + 2 = t$ yields $dx = dt$, $x = t - 2$ and $x^2 + 4x + 6 = t^2 + 2$. (Observe that here $a^2 = 2$.) Then

$$\begin{aligned}\int \frac{2-x}{x^2+4x+6}\,dx &= \int\frac{2-(t-2)}{t^2+2}\,dt\\ &= 4\int\frac{dt}{t^2+2} - \frac{1}{2}\int\frac{2t\,dt}{t^2+2}\\ &= \frac{1}{\sqrt{2}}\tan^{-1}\frac{1}{\sqrt{2}} - \frac{1}{2}\ln(t^2+2) + C\\ &= 2\sqrt{2}\tan^{-1}\frac{x+2}{\sqrt{2}} - \frac{1}{2}\ln(x^2+4x+6) + C. \blacktriangleright\end{aligned}$$

To compute the integral of a partial fraction of the form $\frac{Mx+N}{(x^2+px+q)^k}$, $(k \geqslant 2)$ where the denominator is irreducible to a product of monic linear polynomials we proceed as before by putting $x + \frac{p}{2} = t$ so that $dx = dt$ and $a = \sqrt{q - \frac{p^2}{4}}$. Then

$$\int \frac{Mx+N}{(x^2+px+q)^k}\,dx = \int \frac{Mt+\left(N-\frac{Mp}{2}\right)}{(t^2+a^2)^k}\,dt$$
$$= \frac{M}{2}\int (t^2+a^2)^{-k}\,d(t^2+a^2) + \left(N-\frac{Mp}{2}\right)\int \frac{dt}{(t^2+a^2)^k}$$
$$= \frac{M}{2(1-k)(t^2+a^2)^{k-1}} + \left(N-\frac{Mp}{2}\right)\int \frac{dt}{(t^2+a^2)^k}\,.$$

We denote the integral on the right by J_k and express it as

$$J_k = \int \frac{dt}{(t^2+a^2)^k} = \frac{1}{a^2}\int \frac{(t^2+a^2)-t^2}{(t^2+a^2)^k}\,dt$$
$$= \frac{1}{a^2}\int \frac{dt}{(t^2+a^2)^{k-1}} - \frac{1}{a^2}\int \frac{t^2\,dt}{(t^2+a^2)^k}$$
$$= \frac{1}{a^2}J_{k-1} - \frac{1}{2a^2}\int t\,\frac{d(t^2+a^2)}{(t^2+a^2)^k}\,.$$

To compute the right-hand integral we apply integration by parts by putting $u = t$ and $dv = \frac{d(t^2+a^2)}{(t^2+a^2)^k}$ so that $du = dt$ and $v = \int \frac{d(t^2+a^2)}{(t^2+a^2)^k} = \frac{1}{(1-k)(t^2+a^2)^{k-1}}$. Then

$$J_k = \frac{1}{a^2}J_{k-1} - \frac{1}{2a^2}\left[\frac{t}{(1-k)(t^2+a^2)^{k-1}} - \frac{1}{1-k}\int \frac{dt}{(t^2+a^2)^{k-1}}\right]$$
$$= \frac{1}{a^2}J_{k-1} - \frac{1}{2a^2}\left[\frac{t}{(1-k)(t^2+a^2)^{k-1}} + \frac{1}{k-1}J_{k-1}\right]$$

or

$$J_k = \frac{t}{2a^2(k-1)(t^2+a^2)^{k-1}} + \frac{2k-3}{2a^2(k-1)}J_{k-1}.$$

Thus we have arrived at the *recurrence relation* which yields the value of J_k for any k ($k = 2, 3, \ldots$). Indeed, the integral J_1 is easily computed by using a standard integration formula, i.e.,

$$J_1 = \int \frac{dt}{t^2+a^2} = \frac{1}{a}\tan^{-1}\frac{t}{a} + C.$$

For $k = 2$ the recurrence relation gives

$$J_2 = \int \frac{dt}{(t^2+a^2)^2} = \frac{t}{2a^2(t^2+a^2)} + \frac{J_1}{2a^2}$$
$$= \frac{t}{2a^2(t^2+a^2)} + \frac{1}{2a^3}\tan^{-1}\frac{t}{a} + C.$$

Then we can easily compute J_3 by putting $k = 3$ in the recurrence relation; continuing this process we can find the value of J for any preassigned k. Substituting the relations for t and a into the expression obtained for a given k we finally represent the result in terms of x and M, N, p and q involved in the original integral.

Example. Compute the integral $\int \frac{x+1}{(x^2-4x+5)^2}\,dx$.

◄ Clearly, the integrand is a partial fraction similar to that considered above since the denominator has no real roots due to the fact that $\frac{p^2}{4} - q = -1 < 0$ and, hence, is irreducible to a product of monic linear polynomials and the nominator is a linear polynomial. Then completing the square, we have

$$x^2 - 4x + 5 = (x^2 - 4x + 4) + 1 = (x-2)^2 + 1.$$

The substitution $x - 2 = t$ $(a^2 = 1)$ yields $dx = dt$ and $x = t + 2$ and

$$\int \frac{x+1}{(x^2-4x+5)^2}\,dx = \int \frac{t+3}{(t^2+1)^2}\,dt = \frac{1}{2}\int \frac{2t\,dt}{(t^2+1)^2}$$
$$+3\int \frac{dt}{(t^2+1)^2} = -\frac{1}{2(t^2+1)} + 3\int \frac{dt}{(t^2+1)^2}.$$

Put $k = 2$. Then

$$\int \frac{dt}{(t^2+1)^2} = \frac{t}{2(t^2+1)} + \frac{1}{2}\int \frac{dt}{t^2+1} = \frac{t}{2(t^2+1)} + \frac{1}{2}\tan^{-1} t + C.$$

Hence

$$\int \frac{x+1}{(x^2-4x+5)^2}\,dx = -\frac{1}{2(t^2+1)} + \frac{3t}{2(t^2+1)}$$
$$+\frac{3}{2}\tan^{-1} t + C = \frac{3t-1}{2(t^2+1)} + \frac{3}{2}\tan^{-1} t + C.$$

Finally, the substitution $x = t + 2$ gives-

$$\int \frac{x+1}{(x^2-4x+5)^2}\,dx = \frac{3x-7}{2(x^2-4x+5)} + \frac{3}{2}\tan^{-1}(x-2) + C. \blacktriangleright$$

To summarize the results outlined in the preceding discussion we quote the following theorem.

Theorem 9.4. *The indefinite integral of any rational function always exists on intervals where $Q_n(x) \neq 0$ and is expressed as a composition of a finite number of elementary functions, namely, as an algebraic sum that* *can only involve polynomials, proper rational functions, logarithmic functions or arctangents.*

It is instructive to finalize the general procedure of integrating rational functions as the following sequence of steps: (1) express a given rational function (an integrand) if necessary as a sum of a polynomial and a proper rational function by applying the division algorithm for polynomials or any other suitable technique; (2) factor the denominator of the proper rational function into linear and quadratic polynomials; (3) split the proper rational function into partial fractions; (4) compute the integral of the given functions as a sum of integrals of the summands obtained.

In conclusion of this section we shall consider a number of examples.

Examples. Compute each of the following integrals.

(1) $\displaystyle\int \frac{x^3}{(x-1)(x^2-4)}\,dx.$

◄ Since the integrand $(x-1)(x^2-4) = x^3 - x^2 - 4x + 4$ is not a proper rational function we apply the division algorithm for polynomials

$$\begin{array}{r l}
 & \quad 1 \\
x^3 - x^2 - 4x + 4 & \overline{\big)\, x^3 + 0x^2 + 0x + 0} \\
 & \underline{\;\; x^3 - 1x^2 - 4x - 4} \\
 & \qquad x^2 + 4x - 4.
\end{array}$$

Hence, the integrand becomes

$$\frac{x^3}{(x-1)(x^2-4)} = 1 + \frac{x^2+4x-4}{(x-1)(x^2-4)},$$

where $R_0(x) = 1$ and $\tilde{P}(x) = x^2 + 4x - 4$.

The denominator of the proper rational function has three distinct roots $a = 1$, $b = 2$ and $c = -2$, then this function can be splitted into partial fractions as

$$\frac{x^2+4x-4}{(x-1)(x^2-4)} = \frac{A}{x-1} + \frac{B}{x-2} + \frac{C}{x+2}.$$

Whence

$$x^2 + 4x - 4 = A(x^2-4) + B(x-1)(x+2) + C(x-1)(x-2).$$

Setting x equal to the values of the roots, we obtain

$$A = -1/3 \text{ for } x = 1,\ B = 2 \text{ for } x = 2 \text{ and } C = -2/3 \text{ for } x = -2.$$

Then

$$\frac{x^2+4x-4}{(x-1)(x^2-4)} = -\frac{1}{3}\,\frac{1}{x-1} + 2\,\frac{1}{x-2} - \frac{2}{3}\,\frac{1}{x+2}$$

and

$$\int \frac{x^3\,dx}{(x-1)(x^2-4)} = \int \left(1 - \frac{1}{3}\cdot\frac{1}{x-1} + 2\cdot\frac{1}{x-2} - \frac{2}{3}\cdot\frac{1}{x+2}\right)dx$$

$$= x - \frac{1}{3}\ln|x-1| + 2\ln|x-2| - \frac{2}{3}\ln|x+2| + C. ►$$

(2) $\displaystyle\int \frac{x^2+1}{x^4-x^3}\,dx.$

◄ The integrand is a proper rational function whose denominator has distinct roots $x = 0$ of multiplicity 1 and $x = 1$ of multiplicity 3. Then splitting into partial fractions yields

$$\frac{x^2+1}{x^4-x^3} = \frac{A_1}{x} + \frac{A_2}{x^2} + \frac{A_3}{x^3} + \frac{B}{x-1}.$$

Multiplying out and dividing by the denominator of the integrand, we arrive at

$$x^2 + 1 = A_1x^2(x-1) + A_2(x-1)x + A_3(x-1) + Bx^3$$

or

$$x^2 + 1 = (A_1 + B)x^3 + (-A_1 + A_2)x^2 + (-A_2 + A_3)x - A_3.$$

Comparing the coefficients of x of equal powers, we get $A_1 + B = 0$, $-A_1 + A_2 = 1$, $-A_2 + A_3 = 0$ and $-A_3 = 1$. Whence $A_1 = -2$, $A_2 = -1$, $A_3 = -1$ and $B = 2$. Then

$$\frac{x^2+1}{x^4-x^3} = -\frac{2}{x} - \frac{1}{x^2} - \frac{1}{x^3} + \frac{2}{x-1}$$

and

$$\int \frac{x^2+1}{x^4-x^3}\,dx = \int\left(-\frac{2}{x} - \frac{1}{x^2} - \frac{1}{x^3} + \frac{2}{x-1}\right)dx$$

$$= -2\ln|x| + \frac{1}{x} + \frac{1}{2x^2} + 2\ln|x-1| + C. ►$$

(3) $\displaystyle\int \frac{x^3-x}{(x^2+1)^2}\,dx.$

◄ Since the denominator has no real roots the integrand can be written as

$$\frac{x^3-x}{(x^2+1)^2} = \frac{M_1x+N_1}{x^2+1} + \frac{M_2x+N_2}{(x^2+1)^2}.$$

Then

$$x^3 - x = (M_1x + N_1)(x^2+1) + M_2x + N_2$$

or

$$x^3 - x = M_1x^3 + N_1x^2 + (M_1 + M_2)x + (N_1 + N_2).$$

Comparing the coefficients, we have

$$M_1 = 1, \quad N_1 = 0, \quad M_1 + M_2 = -1, \quad N_1 + N_2 = 0$$

so that

$$M_1 = 1, \quad N_1 = 0, \quad M_2 = -2, \quad N_2 = 0$$

and

$$\int \frac{x^3 - x}{(x^2 + 1)^2}\,dx = \int \left[\frac{x}{x^2 + 1} - \frac{2x}{(x^2 + 1)^2}\right]dx$$

$$= \frac{1}{2}\ln(x^2 + 1) + \frac{1}{x^2 + 1} + C.$$

Observe that the integrand can easily be splitted into partial fractions by using simple algebra as

$$\frac{x^3 - x}{(x^2 + 1)^2} = \frac{(x^3 + x) - 2x}{(x^2 + 1)^2} = \frac{x(x^2 + 1) - 2x}{(x^2 + 1)^2}$$

$$= \frac{x}{x^2 + 1} - \frac{2x}{(x^2 + 1)^2}. \blacktriangleright$$

9.4 Integrals Involving Irrational Functions

Rational and irrational functions. We consider functions in many variables $u_1, u_2, \ldots, u_k$. Let a function $R(u_1, u_2, \ldots, u_k)$ be represented in the form

$$R(u_1, u_2, \ldots, u_k) = \frac{P_m(u_1, u_2, \ldots, u_k)}{Q_n(u_1, u_2, \ldots, u_k)},$$

where $P_m(u_1, u_2, \ldots, u_k)$ and $Q_n(u_1, u_2, \ldots, u_k)$ are the mth and nth degree polynomials in $u_1, u_2, \ldots, u_k$, respectively. Then $R(u_1, u_2, \ldots, u_k)$ is a rational function in $u_1, u_2, \ldots, u_k$; otherwise, it is *irrational.* For example, a quadratic polynomial in variables u_1 and u_2 takes the form $P_1(u_1, u_2) = A_{00} + A_{10}u_1 + A_{01}u_2 + A_{20}u_1^2 + A_{11}u_1u_2 + A_{02}u_2^2$ where the coefficients $A_{00}, A_{10}, A_{01}, A_{20}, A_{11}, A_{02}$ are real numbers and $A_{20}^2 + A_{11}^2 + A_{02}^2 \neq 0$.

It is easy to observe that $f(x, y) = \dfrac{x^2 + 2y^3 + xy}{x + x^3y^2 + 1}$ is a rational function in variables x and y since this function is a ratio of the third degree polynomial $P_3(x, y) = x^2 + 2y^3 + xy$ and the fifth degree polynomial $Q_5(x, y) = x + x^3y^2 + 1$, while $f(x, y) = \dfrac{\sqrt{x^2 - 2xy + 3}}{x + y}$ is an irrational function.

Suppose that variables $u_1, u_2, \ldots, u_k$ are some functions of a variable x, i.e., $u_1 = f_1(x)$, $u_2 = f_2(x)$, $\ldots$, $u_k = f_k(x)$. Then the function $R\,[f_1(x), f_2(x), \ldots, f_k(x)]$ is a rational function in functions $f_1(x), f_2(x), \ldots, f_k(x)$. For example, $f(x) = \dfrac{x^2 + \sqrt{x^2 + x + 1}}{x + 1 + 3\sqrt{x^2 + x + 1}}$ is a rational function in x and in $\sqrt{x^2+x+1}$ so that $f(x) = R\left(x, \sqrt{x^2+x+1}\right)$. It is worth noting that $f(x) = \dfrac{\ln x + e^{\sqrt{x^2+1}}}{2 + \sin x^2}$ is an irrational function in x and in $\sqrt{x^2 + 1}$ while it is a rational function in functions $\ln x$, $e^{\sqrt{x^2+1}}$ and $\sin x^2$ so that $f(x) = R(\ln x, e^{\sqrt{x^2+1}}, \sin x^2)$.

It is not hard to notice that sometimes integrals involving irrational functions do not admit representations in elementary functions. For example, the integrals

$$\int \frac{dx}{\sqrt{(1-x^2)(1-k^2x^2)}} \quad \text{and} \quad \int \frac{x^2\,dx}{\sqrt{(1-x^2)(1-k^2x^2)}} \qquad (0<k<1)$$

called the *elliptic integrals of the first and second kinds*, respectively, can not be expressed in elementary functions. However, by suitable substitutions it is often possible to reduce integrals involving irrational functions to those of rational functions. Below we shall consider techniques applicable in dealing with integrals of some specific forms.

Integrals of the form $\int R\left(x, \sqrt[m]{\dfrac{ax + b}{cx + d}}\right) dx$ where $m \geqslant 2, R(x, y)$ is a rational function in x and $y = \sqrt[m]{\dfrac{ax + b}{cx + d}}$ and the coefficients a, b, c and d are real numbers such that $ad - bc \neq 0$; we leave aside the case when $ad - bc = 0$, for then $\dfrac{a}{c} = \dfrac{b}{d}$ and $\dfrac{ax + b}{cx + d} = \dfrac{a}{c} = \dfrac{b}{d}$ so that the integrand becomes a rational function only in x and its integrals have already been studied in detail.

The substitution

$$t = \sqrt[m]{\frac{ax + b}{cx + d}}$$

yields

$$t^m = \frac{ax + b}{cx + d},$$

$$(cx + d)t^m = ax + b \quad \text{and} \quad ax - cxt^m = dt^m - b$$

so that

$$x = \frac{dt^m - b}{a - ct^m}$$

is a rational function in t.

On differentiating x with respect to t, we have

$$dx = \frac{dmt^{m-1}(a - ct^m) + cmt^{m-1}(dt^m - b)}{(a - ct^m)^2}\, dt$$

and further

$$dx = \frac{(ad - bc)mt^{m-1}}{(a - ct^m)^2}\, dt.$$

Then

$$\int R\left(x, \sqrt[m]{\frac{ax + b}{cx + d}}\right) dx$$

$$= \int R\left(\frac{dt^m - b}{a - ct^m}, t\right) \frac{mt^{m-1}(ad - bc)}{(a - ct^m)^2}\, dt = \int R_1(t)dt,$$

where $R_1(t)$ is a rational function in t the factor $R\left(\frac{dt^m - b}{a - ct^m}, t\right)$ being a function in a rational function is again a rational function and hence the product of this factor by the second factor which is a rational function in t is again a rational function in t.

Therefore we have reduced the original integral to that involving rational functions which are well familiar to us.

Let

$$\int R_1(t)dt = F(t) + C \quad \text{and} \quad F'(t) = R_1(t).$$

Then the original integral becomes

$$\int R\left(x, \sqrt[m]{\frac{ax + b}{cx + d}}\right) dx = F\left(\sqrt[m]{\frac{ax + b}{cx + d}}\right) + C.$$

By way of illustration we consider two examples.

Examples. Compute each of the following integrals.

(1) $\displaystyle\int \sqrt[4]{\frac{2x - 3}{2x + 3}}\, \frac{dx}{(2x + 3)^2}$.

◄ Observe that the integrand is a rational function in x and in $y = \sqrt[4]{\frac{2x - 3}{2x + 3}}$, i.e., $R(x) = \frac{y}{(2x + 3)^2}$.

Put $t = \sqrt[4]{\dfrac{2x-3}{2x+3}}$. Then $2x - 3 = 2xt^4 + 3t^4$, $x = \dfrac{3}{2}\,\dfrac{1+t^4}{1-t^4} = \dfrac{3}{2}\left(\dfrac{2}{1-t^4} - 1\right)$, $dx = \dfrac{12t^3\,dt}{(1-t^4)^2}$ and $2x + 3 = \dfrac{6}{1-t^4}$. Hence the original integral becomes

$$\int \sqrt[4]{\frac{2x-3}{2x+3}}\,\frac{dx}{(2x+3)^2} = \int t\,\frac{(1-t^4)^2}{36}\,\frac{12t^3}{(1-t^4)^2}\,dt$$

$$= \frac{1}{3}\int t^4 dt = \frac{1}{15}t^5 + C = \frac{1}{15}\left(\sqrt[4]{\frac{2x-3}{2x+3}}\right)^5 + C. \blacktriangleright$$

(2) $\displaystyle\int \frac{dx}{\sqrt[4]{x}(\sqrt[3]{x} + \sqrt{x})}$.

◄ Evidently, the integrand is of the form $R(x, \sqrt[12]{x})$. Put $t = \sqrt[12]{x}$. Then $x = t^{12}$, $\sqrt[4]{x} = t^3$, $\sqrt[3]{x} = t^4$, $\sqrt{x} = t^6$ and $dx = 12t^{11}dt$. Hence

$$\int \frac{dx}{\sqrt[4]{x}(\sqrt[3]{x} + \sqrt{x})} = \int \frac{12t^{11}dt}{t^3(t^4 + t^6)} = 12\int \frac{t^4 dt}{1+t^2}$$

$$= 12\int \frac{(t^4 - 1) + 1}{1+t^2}\,dt = 12\int\left(t^2 - 1 + \frac{1}{1+t^2}\right)dt$$

$$= 12\left(\frac{t^3}{3} - t + \tan^{-1} t\right) + C$$

$$= 4\sqrt[3]{x} - 12\sqrt[12]{x} + 12\tan^{-1}\sqrt[12]{x} + C. \blacktriangleright$$

Integrals of the form $\int R\left(x, \sqrt{ax^2 + bx + c}\right) dx$ where the integrand expressed as $R(x, y)$ is a rational function in x and in $y = \sqrt{ax^2 + bx + c}$.

By suitable substitutions dependent on the coefficients a, b and c these integrals are reduced to those involving elementary functions. Three cases are to be distinguished.

(i) Let $a > 0$.

Then the substitution

$$t = \sqrt{ax^2 + bx + c} + \sqrt{a}x$$

yields

$$(t - \sqrt{a}x)^2 = ax^2 + bx + c \quad\text{and}\quad t^2 - 2\sqrt{a}xt = bx + c,$$

so that

$$x = \frac{t^2 - c}{2\sqrt{a}t + b}$$

is a rational function in t.

On differentiating x with respect to t, we get

$$dx = \frac{2t(2\sqrt{a}t + b) - (t^2 - c)2\sqrt{a}}{(2\sqrt{a}t + b)^2}\,dt = 2\,\frac{\sqrt{a}t^2 + bt + c\sqrt{a}}{(2\sqrt{a}t + b)^2}\,dt.$$

We also have

$$\sqrt{ax^2 + bx + c} = t - \sqrt{a}x = t - \sqrt{a}\,\frac{t^2 - c}{2\sqrt{a}t + b}$$
$$= \frac{\sqrt{a}t^2 + bt + c\sqrt{a}}{2\sqrt{a}t + b}.$$

Thus x, dx and $\sqrt{ax^2 + bx + c}$ are all rational functions in t and we can write

$$\int R\left(x, \sqrt{ax^2 + bx + c}\right)dx = \int R_1(t)dt,$$

where

$$R_1(t) = R\left(\frac{t^2 - c}{2\sqrt{a}t + b}, \frac{\sqrt{a}t^2 + bt + c\sqrt{a}}{2\sqrt{a}t + b}\right) 2\,\frac{\sqrt{a}t^2 + bt + c\sqrt{a}}{(2\sqrt{a}t + b)^2}$$

is a rational function in t.

Observe that the substitution $t = \sqrt{ax^2 + bx + c} - \sqrt{a}x$ also reduces the original integral to that involving rational functions.

Example. Compute the integral

$$\int \frac{dx}{\sqrt{x^2 + \alpha^2}}.$$

◄ Since $a = 1 > 0$ we put $t = \sqrt{x^2 + \alpha^2} + x$. Then $x = \frac{t^2 - \alpha^2}{2t}$, $dx = \frac{t^2 + \alpha^2}{2t^2}dt$ and $\sqrt{x^2 + \alpha^2} = t - \frac{t^2 - \alpha^2}{2t} = \frac{t^2 + \alpha^2}{2t}$ so that

$$\int \frac{dx}{\sqrt{x^2 + \alpha^2}} = \int \frac{2t}{t^2 + \alpha^2}\,\frac{t^2 + \alpha^2}{2t^2}\,dt = \int \frac{dt}{t} = \ln|t| + C$$
$$= \ln\left|x + \sqrt{x^2 + \alpha^2}\right| + C.$$

Noting that $x + \sqrt{x^2 + \alpha^2} > 0$ for all x, we finally arrive at

$$\int \frac{dx}{\sqrt{x^2 + \alpha^2}} = \ln\left|(x + \sqrt{x^2 + \alpha^2})\right| + C. \blacktriangleright$$

Problem. Prove that

$$\int \frac{dx}{\sqrt{x^2 - \alpha^2}} = \ln\left|x + \sqrt{x^2 - \alpha^2}\right| + C, \quad |x| > |\alpha|.$$

(ii) Let the quadratic trinomial $ax^2 + bx + c$ have distinct real roots x_1 and x_2 (let the coefficient b be of any sign).

Then the substitution

$$\sqrt{ax^2 + bx + c} = (x - x_1)t \quad (\text{or } \sqrt{ax^2 + bx + c} = (x - x_2)t)$$

yields

$$a(x - x_1)(x - x_2) = (x - x_1)^2t^2 \quad (\text{or } a(x - x_2) = (x - x_1)t)$$

since $ax^2 + bx + c = (x - x_1)^2t^2$.

Whence

$$x = \frac{x_1t^2 - ax_2}{t^2 - a}, \quad dx = \frac{2a(x_2 - x_1)t}{(t^2 - a)^2}dt$$

and

$$\sqrt{ax^2 + bx + c} = \left(\frac{x_1t^2 - ax_2}{t^2 - a} - x_1\right)t = \frac{a(x_1 - x_2)t}{t^2 - a}.$$

Since x, dx and $\sqrt{ax^2 + bx + c}$ are all rational functions in t the original integral becomes

$$\int R\left(x, \sqrt{ax^2 + bx + c}\right)dx = \int R_1(t)dt,$$

where

$$R_1(t) = R\left(\frac{x_1t^2 - ax_2}{t^2 - a}, \frac{a(x_1 - x_2)t}{t^2 - a}\right)\frac{2a(x_2 - x_1)t}{(t^2 - a)^2}$$

is a rational function in t.

Example. Compute the integral

$$\int \frac{dx}{(x - 2)\sqrt{1 - x^2}}, \quad |x| < 1.$$

◄ Since $1 - x^2$ has two distinct real roots $x_1 = -1$ and $x_2 = 1$ we put

$$\sqrt{1 - x^2} = (1 + x)t \quad (\text{or } \sqrt{1 - t^2} = (1 - x)t).$$

Then

$$1 - x^2 = (1 + x)^2t^2, \; 1 - x = (1 + x)t^2, \; x = \frac{1 - t^2}{1 + t^2},$$

$$x - 2 = -\frac{3t^2 + 1}{t^2 + 1}, \; \sqrt{1 - x^2} = \frac{2t}{1 + t^2}, \; dx = -\frac{4t\,dt}{(1 + t^2)^2}$$

and

$$\int \frac{dx}{(x-2)\sqrt{1-x^2}} = \int \frac{(t^2+1)(t^2+1)4t\,dt}{(3t^2+1)2t(t^2+1)^2} = 2\int \frac{dt}{3t^2+1}$$

$$= \frac{2}{3}\int \frac{dt}{t^2+\frac{1}{3}} = \frac{2}{\sqrt{3}}\tan^{-1}\sqrt{3}t + C$$

$$= \frac{2}{\sqrt{3}}\tan^{-1}\sqrt{3\,\frac{1-x}{1+x}} + C. \blacktriangleright$$

(iii) Let $c > 0$.

Then the substitution $\sqrt{ax^2+bx+c} = xt + \sqrt{c}$ (or $\sqrt{ax^2+bx+c} = xt - \sqrt{c}$) easily reduces the original integral to that involving rational functions.

It is worth noting that it suffices to apply the substitutions in (i) and (ii) to reduce integrals of the form $\int R\left(x, \sqrt{ax^2+bx+c}\right)dx$ to those involving rational functions. Indeed, if $b^2 - 4ac > 0$ then the roots of $ax^2 + bx + c$ are real and the substitution given in (ii) yields the desired result. If $b^2 - 4ac < 0$ then the sign of $ax^2 + bx + c$ coincides with that of a; since $ax^2 + bx + c$ must be positive, then $a > 0$. In this case the substitution shown in (i) leads to what we are seeking for. Notice that some integrals of the form $\int R\left(x, \sqrt{ax^2+bx+c}\right)dx$ can easily be computed without applying the substitutions mentioned in (i)-(iii). Below we shall consider three specific cases.

(a) *Integrals of the form* $\displaystyle\int \frac{dx}{\sqrt{ax^2+bx+c}}$, $a \neq 0$. Here it is suitable to complete the square of the radicand as

$$ax^2+bx+c = a\left(x^2 + 2x\frac{b}{2a} + \frac{c}{a}\right)$$

$$= a\left[\left(x^2 + 2x\frac{b}{2a} + \frac{b^2}{4a^2}\right) + \frac{c}{a} - \frac{b^2}{4a^2}\right]$$

$$= a\left[\left(x + \frac{b}{2a}\right)^2 + \frac{4ac-b^2}{4a^2}\right]$$

$$= a\left(x + \frac{b}{2a}\right)^2 + \frac{4ac-b^2}{4a} = a\left(x + \frac{b}{2a}\right)^2 + p,$$

where $p = \dfrac{4ac-b^2}{4a}$.

Then the substitution $t = x + \frac{b}{2a}$ yields $dx = dt$ and

$$\int \frac{dx}{\sqrt{ax^2 + bx + c}} = \int \frac{dt}{\sqrt{at^2 + p}},$$

where a and p are of opposite signs or both positive.

When $a > 0$ and $p > 0$ and when $a < 0$ and $p < 0$ the integral is expressed by a logarithmic function; when $a < 0$ and $p > 0$ the integral is expressed by an arcsine as the following two examples illustrate.

Examples. (1) $\int \frac{dx}{\sqrt{x^2 - 4x + 5}}$.

◄ Since $x^2 - 4x + 5 = (x - 2)^2 + 1$ then the substitution $x - 2 = t$ gives $dx = dt$ and

$$\int \frac{dx}{\sqrt{x^2 - 4x + 5}} = \int \frac{dx}{\sqrt{(x - 2)^2 + 1}} = \int \frac{dt}{\sqrt{t^2 + 1}}$$

$$= \ln\left(t + \sqrt{t^2 + 1}\right) + C = \ln\left(x - 2 + \sqrt{x^2 - 4x + 5}\right) + C. \text{ ►}$$

(2) $\int \frac{dx}{\sqrt{6x - x^2}}$.

◄ Since $6x - x^2 = -[(x^2 - 6x + 9) - 9] = 9 - (x - 3)^2$ the substitution $x - 3 = t$ yields $dx = dt$ and

$$\int \frac{dx}{\sqrt{6x - x^2}} = \int \frac{dt}{\sqrt{9 - t^2}} = \sin^{-1}\frac{t}{3} + C = \sin^{-1}\frac{x - 3}{3} + C. \text{ ►}$$

(b) *Integrals of the form* $\int \frac{Mx + N}{\sqrt{ax^2 + bx + c}}\,dx$, $a \neq 0$. Noting that $(ax^2 + bx + c)' = 2ax + b$, we reduce the original integral to that considered above by proceeding as

$$\int \frac{Mx + N}{\sqrt{ax^2 + bx + c}}\,dx = \int \frac{M\frac{1}{2a}[(2ax + b) - b] + N}{\sqrt{ax^2 + bx + c}}\,dx$$

$$= \frac{M}{2a}\int \frac{(2ax + b)dx}{\sqrt{ax^2 + bx + c}} + \left(N - \frac{Mb}{2a}\right)\int \frac{dx}{\sqrt{ax^2 + bx + c}}$$

$$= \frac{M}{2a}\int \frac{d(ax^2 + bx + c)}{\sqrt{ax^2 + bx + c}} + \left(N - \frac{Mb}{2a}\right)\int \frac{dx}{\sqrt{ax^2 + bx + c}}$$

$$= \frac{M}{a}\sqrt{ax^2 + bx + c} + \left(N - \frac{Mb}{2a}\right)\int \frac{dx}{\sqrt{ax^2 + bx + c}}.$$

Example. Compute the integral

$$\int \frac{x+1}{\sqrt{6x-x^2}}\,dx.$$

◄ Noting that $(6x - x^2)' = 6 - 2x$, we obtain

$$\int \frac{x+1}{\sqrt{6x-x^2}}\,dx = -\frac{1}{2}\int \frac{-2x-2}{\sqrt{6x-x^2}}\,dx = -\frac{1}{2}\int \frac{(6-2x)-8}{\sqrt{6x-x^2}}\,dx$$

$$=4\int \frac{dx}{\sqrt{6x-x^2}} - \frac{1}{2}\int \frac{d(6x-x^2)}{\sqrt{6x-x^2}} = 4\sin^{-1}\frac{x-3}{3} - \sqrt{6x^2-x^2} + C. ►$$

(c) *Integrals of the form* $\int \frac{P_n(x)}{\sqrt{ax^2+bx+c}}\,dx$ where $P_n(x)$ is a polynomial of degree n. Here we apply the method of comparing coefficients to find the coefficients of $P_n(x)$.

Let there hold the identity

$$\int \frac{P_n(x)}{\sqrt{ax^2+bx+c}}\,dx = Q_{n-1}(x)\sqrt{ax^2+bx+c} + A_n\int \frac{dx}{\sqrt{ax^2+bx+c}}, \qquad (*)$$

where $Q_{n-1}(x)$ is a polynomial of degree $(n-1)$ with unknown coefficients $A_0, A_1, \ldots, A_{n-1}$, i.e.,

$$Q_{n-1}(x) = A_0 + A_1x + \ldots + A_{n-1}x^{n-1}.$$

To compute the values of the unknown coefficients we differentiate the above identity with respect to x so that

$$\frac{P_n(x)}{\sqrt{ax^2+bx+c}} = Q'_{n-1}(x)\sqrt{ax^2+bx+c} + Q_{n-1}(x)\frac{ax+b/2}{\sqrt{ax^2+bx+c}} + \frac{A_n}{\sqrt{ax^2+bx+c}}.$$

Multiplying throughout by $\sqrt{ax^2+bx+c}$, we get the identity

$$P_n(x) = Q'_{n-1}(x)(ax^2+bx+c) + Q_{n-1}(x)\left(ax+\frac{b}{2}\right) + A_n$$

which involves the nth degree polynomials on both sides.

Comparing the coefficients of x of equal powers, we get $(n+1)$ equations which yield the desired coefficients A_k $(k = 0, 1, 2, \ldots, n)$. Then substituting these coefficients into the right-hand side of (*) and computing the integral $\int \frac{dx}{\sqrt{ax^2+bx+c}}$, we arrive at the desired result.

Example. Compute the integral

$$\int \frac{x^2dx}{\sqrt{x^2+2x+2}}.$$

◀ Put

$$\int \frac{x^2dx}{\sqrt{x^2+2x+2}} = (A_0 + A_1x)\sqrt{x^2+2x+2} + A_2\int\frac{dx}{\sqrt{x^2+2x+2}}. \qquad (**)$$

On differentiating both sides, we have

$$\frac{x^2}{\sqrt{x^2+2x+2}} = A_1\sqrt{x^2+2x+2} + (A_0 + A_1x)\frac{x+1}{\sqrt{x^2+2x+2}} + \frac{A_2}{\sqrt{x^2+2x+2}}.$$

Multiplying out by $\sqrt{x^2+2x+2}$, we get

$$x^2 = A_1(x^2+2x+2) + (A_0 + A_1x)(x+1) + A_2$$

or

$$x^2 = 2A_1x^2 + (2A_1 + A_1 + A_0)x + (A_0 + 2A_1 + A_2).$$

The method of comparing coefficients gives

$$\begin{array}{l|l} x^2 & 2A_1 = 1 \\ x^1 & A_0 + 3A_1 = 0 \\ x^0 & A_0 + 2A_1 + A_2 = 0 \end{array} \Bigg\}.$$

Whence

$$A_0 = -\frac{3}{2},\ A_1 = \frac{1}{2},\ A_2 = \frac{1}{2}.$$

Computing the integral in the right-hand side of (**), we have

$$\int\frac{dx}{\sqrt{x^2+2x+2}} = \int\frac{d(x+1)}{\sqrt{(x+1)^2+1}} = \ln\left(x+1+\sqrt{x^2+2x+2}\right) + C.$$

Then the original integral becomes

$$\int\frac{x^2dx}{\sqrt{x^2+2x+2}} = \frac{x-3}{2}\sqrt{x^2+2x+2} + \frac{1}{2}\ln\left(x+1+\sqrt{x^2+2x+2}\right) + C. \ ▶$$

9.5 Integrals Involving Trigonometric Functions

Integrals of the form $\int R(\sin x, \cos x)\,dx$ where the integrand is a rational function in both $\sin x$ and $\cos x$. For example, $f(x) = \dfrac{1 - 2\sin x}{2 + \cos^2 x}$ is a rational function in both $\sin x$ and $\cos x$; $g(x) = \dfrac{1 + \sin^2 x}{\sqrt{\cos x} + \cos x}$ is a rational function in $\sin x$ and is irrational in $\cos x$ (we shall not examine functions of this kind here).

By the substitution $\tan\frac{x}{2} = t$ where $-\pi < x < \pi$, the original integral is reduced to that in a rational function. Indeed,

$$\sin x = \frac{2\sin\frac{x}{2}\cos\frac{x}{2}}{\cos^2\frac{x}{2} + \sin^2\frac{x}{2}} = \frac{2\tan\frac{x}{2}}{1 + \tan^2\frac{x}{2}} = \frac{2t}{1 + t^2},$$

$$\cos x = \frac{\cos^2\frac{x}{2} - \sin^2\frac{x}{2}}{\cos^2\frac{x}{2} + \sin^2\frac{x}{2}} = \frac{1 - \tan^2\frac{x}{2}}{1 + \tan^2\frac{x}{2}} = \frac{1 - t^2}{1 + t^2},$$

$$x = 2\tan^{-1} t, \quad dx = \frac{2dt}{1 + t^2}$$

and

$$\int R(\sin x, \cos x)\,dx = \int R\left(\frac{2t}{1 + t^2}, \frac{1 - t^2}{1 + t^2}\right)\frac{2\,dt}{1 + t^2} = \int R_1(t)\,dt,$$

where $R_1(t)$ is a rational function in t.

Example. Compute the integral $\displaystyle\int \frac{dx}{\sin x}$.

◄ The substitution $\tan\frac{x}{2} = t$ yields

$$\int \frac{dx}{\sin x} = \int \frac{\frac{2\,dt}{1 + t^2}}{\frac{2t}{1 + t^2}} = \int \frac{dt}{t} = \ln|t| + C = \ln\left|\tan\frac{x}{2}\right| + C. ►$$

Sometimes the substitution $\tan\frac{x}{2} = t$ requires unwieldy computations; so we shall mention three specific cases when integrals of the form $\int R(\sin x, \cos x)\,dx$ can be computed by using simpler substitutions.

(a) *Integrals of the form* $\int R(\sin x)\cos x\,dx$. Here the substitution $\sin x = t$ gives $\cos x\,dx = dt$ so that the original integral becomes $\int R(t)\,dt$. For example, if we wish to compute the integral

$$\int \frac{\cos x\,dx}{4+\sin^2 x},$$

then the substitution $\sin x = t$ yields $dt = \cos x\,dx$ and

$$\int \frac{\cos x\,dx}{4+\sin^2 x} = \int \frac{dt}{4+t^2} = \frac{1}{2}\tan^{-1}\frac{t}{2} + C = \frac{1}{2}\tan^{-1}\left(\frac{1}{2}\sin x\right) + C. \blacktriangleright$$

(b) *Integrals of the form* $\int R(\cos x)\sin x\,dx$. The substitution $\cos x = t$ yields $\sin x\,dx = -dt$ so that the original integral becomes $-\int R(t)\,dt$. For example, using this substitution we can easily compute the following integral

$$\int \frac{\sin x}{2+\cos x}\,dx = -\int \frac{dt}{2+t} = -\int \frac{d(2+t)}{2+t}$$
$$= -\ln(2+t) + C = -\ln(2+\cos x) + C.$$

(c) *Integrals of the form* $\int R(\sin x, \cos x)\,dx$ where the integrand $R(\sin x, \cos x)$ involves only even powers of $\sin x$ and $\cos x$. Then the substitution $\tan x = t$ gives $x = \tan^{-1} t$ and $dx = \dfrac{dt}{1+t^2}$ so that $\sin^2 x$ and $\cos^2 x$ being rational functions in $\tan x$ become rational functions in t. Indeed, we have

$$\sin^2 x = \frac{\sin^2 x}{\cos^2 x + \sin^2 x} = \frac{\tan^2 x}{1+\tan^2 x} = \frac{t^2}{1+t^2},$$
$$\cos^2 x = \frac{\cos^2 x}{\cos^2 x + \sin^2 x} = \frac{1}{1+\tan^2 x} = \frac{1}{1+t^2},$$

so that

$$\int R(\sin x, \cos x)\,dx = \int R\left(\frac{t^2}{1+t^2}, \frac{1}{1+t^2}\right)\frac{dt}{1+t^2} = \int R_1(t)\,dt,$$

where $R_1(t)$ is a rational function in t.

To demonstrate how this substitution works we compute

$$\int \frac{dx}{\sin^2 x + 4\cos^2 x + 2}.$$

◄ Put $\tan x = t$. Then $dx = \dfrac{dt}{1+t^2}$, $\sin^2 x = \dfrac{t^2}{1+t^2}$ and $\cos^2 x = \dfrac{1}{1+t^2}$ so that

$$\int \frac{dx}{\sin^2 x + 4\cos^2 x + 2} = \int \frac{1}{\dfrac{t^2}{1+t^2} + \dfrac{4}{1+t^2} + 2} \, \frac{dt}{1+t^2}$$

$$= \frac{1}{3}\int \frac{dt}{t^2+2} = \frac{1}{3\sqrt{2}} \tan^{-1} \frac{t}{\sqrt{2}} + C = \frac{1}{3\sqrt{2}} \tan^{-1}\left(\frac{\tan x}{\sqrt{3}}\right) + C. \text{ ►}$$

Integrals of the form $\int \sin^\alpha x \cos^\alpha x \, dx$ where α and β are real numbers.

We consider two cases when the integral admits a representation in rational functions.

(a) Let either α or β be an odd positive number. For definiteness we put $\beta = 2k + 1$ where $k > 0$ is an integer. Then α is any real number. Using the identity $\cos^2 x + \sin^2 x = 1$, we have

$$\int \sin^\alpha x \cos^\beta x \, dx = \int \sin^\alpha x \cos^{2k+1} x \, dx = \int \sin^\alpha x \, (\cos^2 x)^k \cos x \, dx$$
$$= \int \sin^\alpha x \, (1 - \sin^2 x)^k \cos x \, dx.$$

Put $\sin x = t$. Then $\cos x \, dx = dt$ and

$$\int \sin^\alpha x \cos^{2k+1} x \, dx = \int t^\alpha (1 - t^2)^k \, dt.$$

Applying the binomial theorem to the integrand on the right, we obtain $(k + 1)$ power functions which are easily integrable.

Examples. Compute each of the following integrals.

(1) $\int \sin^2 x \cos^5 x \, dx.$

◄ Manipulating on the integrand, we have

$$\int \sin^2 x \cos^5 x \, dx = \int \sin^2 x \cos^4 x \cos x \, dx$$
$$= \int \sin^2 x (1 - \sin^2 x)^2 \cos x \, dx.$$

Then the substitution $\sin x = t$ gives $\cos x \, dx = dt$ and

$$\int \sin^2 x (1 - \sin^2 x)^2 \cos x \, dx = \int t^2 (1 - t^2)^2 \, dt$$
$$= \int (t^2 - 2t^4 + t^6) \, dt = \frac{1}{3} t^3 - \frac{2}{5} t^5 + \frac{1}{7} t^7 + C$$
$$= \frac{1}{3} \sin^3 x - \frac{2}{5} \sin^5 x + \frac{1}{7} \sin^7 x + C. \text{ ►}$$

(2) $\displaystyle\int \frac{\sin^3 x}{\cos^2 x}\, dx.$

◄ We have

$$\int \frac{\sin^3 x}{\cos^2 x}\, dx = \int \frac{\sin^2 x}{\cos^2 x} \sin x \, dx = \int \frac{1 - \cos^2 x}{\cos^2 x} \sin x \, dx.$$

Put $\cos x = t$. Then $\sin x \, dx = -dt$ and

$$\int \frac{1 - \cos^2 x}{\cos^2 x} \sin x \, dx = \int \frac{t^2 - 1}{t^2}\, dt = \int \left(1 - \frac{1}{t^2}\right) dt$$
$$= t + \frac{1}{t} + C = \cos x + \frac{1}{\cos x} + C. \blacktriangleright$$

(3) $\displaystyle\int \frac{\cos^3 x}{\sqrt{\sin x}}\, dx.$

◄ We have

$$\int \frac{\cos^3 x}{\sqrt{\sin x}}\, dx = \int \frac{\cos^2 x}{\sqrt{\sin x}} \cos x \, dx = \int \frac{1 - \sin^2 x}{\sqrt{\sin x}} \cos x \, dx$$

Put $\sin x = t$. Then $\cos x \, dx = dt$ and

$$\int \frac{1 - \sin^2 x}{\sqrt{\sin x}} \cos x \, dx = \int \frac{1 - t^2}{\sqrt{t}}\, dt = \int (t^{-1/2} - t^{3/2})\, dt$$
$$= \sqrt{2}t^{1/2} - \frac{2}{5} t^{5/2} + C = 2\sqrt{\sin x} - \frac{2}{5} (\sqrt{\sin x})^5 + C. \blacktriangleright$$

(b) Let α and β be positive even numbers, i.e., $\alpha = 2m$ and $\beta = 2n$ where m and n are natural numbers. Here it is helpful to manipulate on the integrand by applying the trigonometric identities

$$\sin^2 x = \frac{1 - \cos 2x}{2} \quad \text{and} \quad \cos^2 x = \frac{1 + \cos 2x}{2}. \tag{$*$}$$

Suppose that $m \neq n$. Then using the identities $(*)$, we obtain

$$\int \sin^{2m} x \cos^{2n} x \, dx = \int (\sin^2 x)^m (\cos^2 x)^n \, dx$$
$$= \int \left(\frac{1 - \cos 2x}{2}\right)^m \left(\frac{1 + \cos 2x}{2}\right)^n dx$$
$$= \frac{1}{2^{m+n}} \int (1 - \cos 2x)^m (1 + \cos 2x)^n \, dx.$$

Applying the binomial theorem to the factors $(1 - \cos 2x)^m$ and $(1 + \cos 2x)^n$ and multiplying the polynomials thus obtained we arrive at the integrand involving odd and even powers of $\cos 2x$. The terms involving odd powers of $\cos 2x$ are easily integrable as we have discussed in the preceding section. To integrate the terms involving even powers of $\cos 2x$ we apply to them the identity (*). This yields the terms involving powers of $\cos 4x$. Continuing this process we finally arrive at integrals of the form $\int \cos kx\,dx$ where $k > 0$ is a positive number. These integrals can be computed without difficulty.

Suppose now that $m = n$. Then applying the identity

$$\sin x \cos x = \frac{1}{2} \sin 2x$$

we obtain

$$\int \sin^{2n} x \cos^{2n} x\,dx = \int (\sin x \cos x)^{2n}\,dx = \int \left(\frac{1}{2} \sin 2x\right)^{2n} dx$$
$$= \frac{1}{4^n} \int \sin^{2n} 2x\,dx = \frac{1}{4^n} \int (\sin^2 2x)^n\,dx$$
$$= \frac{1}{4^n} \int \left(\frac{1 - \cos 4x}{2}\right)^n dx = \frac{1}{8^n} \int (1 - \cos 4x)^n\,dx.$$

Clearly, the integral on the right is easily computed as given above.

Examples. Compute each of the following integrals.

(1) $\int \sin^2 x \cos^4 x\,dx.$

◄ We have

$$\int \sin^2 x \cos^4 x\,dx = \int \frac{1 - \cos 2x}{2} \left(\frac{1 + \cos 2x}{2}\right)^2 dx$$
$$= \frac{1}{8} \int (1 + \cos 2x - \cos^2 2x - \cos^3 2x)\,dx$$
$$= \frac{1}{8} \int \left[1 + \cos 2x - \frac{1 + \cos 4x}{2} - (1 - \sin^2 2x)\cos 2x\right] dx$$
$$= \frac{1}{8} \int \left(\frac{1}{2} - \frac{1}{2} \cos 4x + \sin^2 2x \cos 2x\right) dx$$
$$= \frac{1}{8}\left(\frac{1}{2} x - \frac{1}{8} \sin 4x + \frac{1}{6} \sin^3 2x\right) + C. \blacktriangleright$$

(2) $\int \sin^2 x \cos^2 x\,dx.$

◄ We have

$$\int \sin^2 x \cos^2 x\,dx = \int (\sin x \cos x)^2\,dx = \frac{1}{4} \int \sin^2 2x\,dx$$
$$= \frac{1}{4} \int \frac{1 - \cos 4x}{2}\,dx = \frac{1}{8}\left(x - \frac{1}{4} \sin 4x\right) + C. \blacktriangleright$$

Integrals of the form $\int \sin \alpha x \cos \beta x\, dx$, $\int \cos \alpha x \cos \beta x\, dx$, $\int \sin \alpha x \sin \beta x\, dx$ where $\alpha \neq \beta$. To compute these integrals it is helpful to use the following trigonometric identities

$$\sin \alpha x \cos \beta x = \frac{1}{2}\,[\sin(\alpha + \beta)x + \sin(\alpha - \beta)x],$$

$$\cos \alpha x \cos \beta x = \frac{1}{2}\,[\cos(\alpha + \beta)x + \cos(\alpha - \beta)x],$$

$$\sin \alpha x \sin \beta x = \frac{1}{2}\,[\cos(\alpha - \beta)x - \cos(\alpha + \beta)x].$$

For example, if we wish to compute $\int \sin \alpha x \cos \beta x\, dx$ then

$$\int \sin \alpha x \cos \beta x\, dx = \frac{1}{2}\int [\sin(\alpha + \beta)x + \sin(\alpha - \beta)x]\, dx$$

$$= \frac{1}{2}\left[-\frac{\cos(\alpha + \beta)x}{\alpha + \beta} - \frac{\cos(\alpha - \beta)x}{\alpha - \beta}\right] + C.$$

The other two integrals are computed in a similar way.

Example. Compute $\int \cos 3x \cos x\, dx$.

◄ We have

$$\int \cos 3x \cos x\, dx = \frac{1}{2}\int (\cos 4x + \cos 2x)\, dx$$

$$= \frac{1}{2}\left(\frac{1}{4}\sin 4x + \frac{1}{2}\sin 2x\right) + C = \frac{1}{8}\sin 4x + \frac{1}{4}\sin 2x + C.$$ ►

Exercises

Apply the standard integration formulas to compute the following integrals

1. $\int x^2 \sqrt[3]{x}\, dx$. **2.** $\int \frac{dx}{\sqrt[4]{x^3}}$. **3.** $\int \sqrt{x\sqrt{x\sqrt{x}}}\, dx$. **4.** $\int \left(\sqrt{x} + \frac{1}{\sqrt{x}}\right)^2 dx$.

5. $\int 2^x 8^x\, dx$. **6.** $\int \frac{6^x}{3^{2x}}\, dx$. **7.** $\int \frac{2^x + 5^x}{10^x}\, dx$. **8.** $\int \frac{\cos 2x}{\cos^2 x \sin^2 x}\, dx$.

9. $\int \frac{\sin 2x}{\sin x \cos^3 x}\, dx$. **10.** $\int \frac{\sin 7x + \sin 3x}{\sin 5x \cos 2x}\, dx$. **11.** $\int (\tan x - \cot x)^2 dx$.

12. $\int \frac{dx}{\cos^2 x \sin^2 x}$. **13.** $\int \frac{dx}{\sinh^2 x \cosh^2 x}$. **14.** $\int \frac{\sinh 2x}{\cosh x}\, dx$.

15. $\int \tanh^2 x\,dx$. **16.** $\int \coth^2 x\,dx$. **17.** $\int \frac{dx}{4x^2+9}$. **18.** $\int \frac{dx}{4x^2-9}$.

19. $\int \frac{dx}{\sqrt{4x^2+9}}$. **20.** $\int \frac{dx}{\sqrt{9-4x^2}}$.

Use integration by substitution to compute the following integrals

21. $\int xe^{x^2/2}\,dx$. **22.** $\int x^2 \sin \frac{x^3}{3}\,dx$. **23.** $\int \frac{x}{\sqrt{1-x^2}}\,dx$. **24.** $\int \frac{dx}{x \ln x}$.

25. $\int \frac{dx}{\sqrt[4]{x^3}\,(1+\sqrt[4]{x})}$. **26.** $\int \frac{dx}{\sqrt[4]{x^3}\sqrt{1-\sqrt[4]{x}}}$. **27.** $\int \frac{dx}{\sqrt{e^x+1}}$.

28. $\int \frac{e^{2x}}{e^x-1}\,dx$. **29.** $\int x^2(x-1)^{18}\,dx$. **30.** $\int \frac{dx}{x\sqrt{x-1}}$. **31.** $\int x\sqrt[3]{x+1}\,dx$.

32. $\int \frac{x}{\sqrt[3]{x+1}}\,dx$. **33.** $\int \frac{x^3}{\sqrt{1-x^2}}\,dx$. **34.** $\int \frac{x\,dx}{1+x^4}$. **35.** $\int \frac{x^2}{\sqrt{1-x^6}}\,dx$.

36. $\int \frac{\ln \tan x}{\sin 2x}\,dx$. **37.** $\int \frac{\cot x}{\ln \sin x}\,dx$. **38.** $\int \frac{\sin x}{\cos^3 x}\,e^{\tan^2 x}\,dx$.

39. $\int \frac{\tan^{-1} x\, e^{(\tan^{-1})^2 x}}{1+x^2}\,dx$. **40.** $\int \frac{\ln x}{x(4+\ln^2 x)}\,dx$.

Use integration by parts to compute the following integrals

41. $\int xe^{-x}\,dx$. **42.** $\int x2^x\,dx$. **43.** $\int x \sin 2x\,dx$. **44.** $\int (1+x)e^x\,dx$.

45. $\int (1+x\ln 2)2^x\,dx$. **46.** $\int (2x-x^2)e^{-x}\,dx$. **47.** $\int \tan^{-1} x\,dx$.

48. $\int x\tan^{-1} x\,dx$. **49.** $\int \sin^{-1} x\,dx$. **50.** $\int \frac{x}{\cos^2 x}\,dx$. **51.** $\int x\tan^2 x\,dx$.

52. $\int \cos(\ln x)\,dx$. **53.** $\int \sin(\ln x)\,dx$.

Integrate partial fractions

54. $\int \frac{2\,dx}{5+2x}$. **55.** $\int \frac{dx}{2-3x}$. **56.** $\int \frac{dx}{(3x+5)^3}$. **57.** $\int \frac{dx}{(1-2x)^5}$.

58. $\int \frac{dx}{x^2+2x+3}$. **59.** $\int \frac{x+2}{x^2+2x+5}\,dx$. **60.** $\int \frac{x+1}{x^2-x+2}\,dx$.

61. $\int \frac{x\,dx}{x^2+7x+13}$.

Applying the method of comparing coefficients, compute the following integrals

62. $\int \frac{2x+3}{x^2+3x-10}\,dx$. **63.** $\int \frac{x\,dx}{(x+1)(2x+1)}$. **64.** $\int \frac{3x^2-2x-4}{(x-1)(x^2-4)}\,dx$.

65. $\int \frac{x^3-x+2}{x^2-1}\,dx$. **66.** $\int \frac{x^4+1}{x^3-x}\,dx$. **67.** $\int \frac{x+2}{(x+3)^2}\,dx$. **68.** $\int \frac{x^2+2x}{(x+1)^4}\,dx$.

69. $\int \frac{dx}{x^3+x^4}$. **70.** $\int \frac{x^2-3x}{(x+1)(x-1)^2}\,dx$. **71.** $\int \frac{dx}{x(x^2+1)}$. **72.** $\int \frac{dx}{x^4-1}$.

73. $\int \frac{x\,dx}{x^3-1}$.

Compute the integrals involving irrational functions

74. $\int \frac{1}{(1-x)^2}\sqrt{\frac{1-x}{1+x}}\,dx$. **75.** $\int \frac{\sqrt{1+x}}{x}\,dx$.

76. $\int \frac{1}{(1-x)(1+x)^2}\sqrt{\frac{1+x}{1-x}}\,dx$. **77.** $\int \sqrt[3]{\frac{1-x}{1+x}}\,\frac{dx}{(1+x)^2}$.

78. $\int \frac{dx}{\sqrt{x^2-4x}}$. **79.** $\int \frac{dx}{\sqrt{4x-x^2}}$. **80.** $\int \frac{dx}{\sqrt{x^2-4x-5}}$.

81. $\int \frac{dx}{\sqrt{x^2-4x+5}}$. **82.** $\int \frac{dx}{\sqrt{5-4x-x^2}}$. **83.** $\int \frac{x+1}{\sqrt{x^2-6x-1}}\,dx$.

84. $\int \frac{x+1}{\sqrt{1+6x-x^2}}\,dx$. **85.** $\int \frac{x-3}{\sqrt{1+6x-x^2}}\,dx$. **86.** $\int \frac{2x^2+3x+2}{\sqrt{x^2+2x+2}}\,dx$.

87. $\int \frac{1+2x-3x^3}{\sqrt{1-x^2}}\,dx$. **88.** $\int \frac{2x^2+1}{\sqrt{x^2+2}}\,dx$. **89.** $\int \frac{x^5\,dx}{\sqrt{1-x^2}}$. **90.** $\int \frac{x^6}{\sqrt{1+x^2}}\,dx$.

Compute the integrals involving trigonometric functions

91. $\int \frac{dx}{2+\cos x}$. **92.** $\int \frac{dx}{5+4\sin x}$. **93.** $\int \frac{dx}{3\sin x-4\cos x}$.

94. $\int \frac{dx}{5+\sin x+3\cos x}$. **95.** $\int \frac{\sin 2x}{1+\sin^2 x}\,dx$. **96.** $\int \frac{\cos x\,dx}{5+\sin^2 x-6\sin x}$.

97. $\int \frac{1+\cos x-\sin x}{1-\cos x+\sin x}\,dx$. **98.** $\int \cos^3 x\,dx$. **99.** $\int \sin^5 x\,dx$.

100. $\int \sin^2 x \cos^3 x\, dx.$ **101.** $\int \cos^4 x \sin^3 x\, dx.$ **102.** $\int \frac{\sin^3 x}{\cos^4 x}\, dx.$

103. $\int \sin^2 x \cos^2 x\, dx.$ **104.** $\int \sin^4 x\, dx.$ **105.** $\int \sin^4 x \cos^2 x\, dx.$

106. $\int \frac{dx}{\sin^4 x}.$ **107.** $\int \frac{\cos^2 x}{\sin^4 x}\, dx.$ **108.** $\int \frac{dx}{\sin^2 x \cos^4 x}.$

109. $\int \frac{dx}{\cos^4 x - \sin^4 x}.$ **110.** $\int \sin 5x \cos x\, dx.$ **111.** $\int \sin x \cos 5x\, dx.$

112. $\int \cos 7x \cos 3x\, dx.$ **113.** $\int \sin 15x \sin 10x\, dx.$ **114.** $\int \cos \frac{x}{2} \cos \frac{x}{3}\, dx.$

115. $\int \sin x \sin 2x \sin 3x\, dx.$

Answers

1. $0.3\, x^{10/3} + C.$ **2.** $4\sqrt[4]{x} + C.$ **3.** $\frac{8}{15} x^{15/8} + C.$ **4.** $\frac{x^2}{2} + 2x + \ln|x| + C.$

5. $\frac{16^x}{\ln 16} + C.$ **6.** $\frac{1}{\ln \frac{2}{3}} \left(\frac{2}{3}\right)^x + C.$ **7.** $-\frac{5^{-x}}{\ln 5} - \frac{2^{-x}}{\ln 2} + C.$ **8.** $-\cot x - \tan x + C.$

9. $2\tan x + C.$ **10.** $2x + C.$ **11.** $\tan x - \cot x - 4x + C.$ **12.** $\tan x - \cot x + C.$

13. $-\tanh x - \coth x + C.$ **14.** $2\cosh x + C.$ **15.** $x - \tanh x + C.$ **16.** $x - \coth x + C.$

17. $\frac{1}{6} \tan^{-1} \frac{2}{3} x + C.$ **18.** $\frac{1}{12} \ln \left| \frac{2x - 3}{2x + 3} \right| + C.$ **19.** $\ln \sqrt{2x + \sqrt{4x^2 + 9}} + C.$

20. $\frac{1}{2} \sin^{-1} \frac{2}{3} x + C.$ **21.** $e^{x^2/2} + C.$ **22.** $-\cos \frac{x^3}{3} + C.$ **23.** $-\sqrt{1 - x^2} + C.$ **24.** $\ln|\ln x| + C.$

25. $4 \ln (1 + \sqrt[4]{x}) + C.$ **26.** $-8\sqrt{1 - \sqrt[4]{x}} + C.$ **27.** $\ln \frac{\sqrt{e^x + 1} - 1}{\sqrt{e^x + 1} + 1} + C.$ **28.** $\ln|e^x - 1| +$
$e^x + C.$ **29.** $\frac{(x - 1)^{21}}{21} + \frac{(x - 1)^{20}}{10} + \frac{(x - 1)^{19}}{19} + C.$ **30.** $2\tan^{-1} \sqrt{x - 1}.$

31. $\frac{3}{28} (x + 1)^{4/3} (4x - 3) + C.$ **32.** $\frac{3}{10} (x + 1)^{2/3} (2x - 3) + C.$ **33.** $-(2 + x^2)\sqrt{1 - x^2} + C.$

34. $\frac{1}{2} \tan^{-1} x^2 + C.$ **35.** $\frac{1}{3} \sin^{-1} x^3 + C.$ **36.** $\frac{1}{4} \ln^2 \tan x + C.$ **37.** $\ln|\ln \sin x| + C.$ **38.** $\frac{1}{2} e^{\tan^2 x} + C.$

39. $\frac{1}{2} e^{(\tan^{-1})^2 x} + C.$ **40.** $\ln \sqrt{4 + \ln^2 x} + C.$ **41.** $-(x + 1)e^{-x} + C.$ **42.** $\frac{x \ln 2 - 1}{\ln^2 2} 2^x + C.$

43. $\frac{1}{4} \sin 2x - \frac{x}{2} \cos 2x + C.$ **44.** $xe^x + C.$ **45.** $x\, 2^x + C.$ **46.** $x^2 e^{-x} + C.$ **47.** $x \tan^{-1} x -$
$\ln \sqrt{1 + x^2} + C.$ **48.** $\frac{1}{2} (x^2 + 1) \tan^{-1} x - \frac{1}{2} x + C.$ **49.** $x \sin^{-1} x + \sqrt{1 - x^2} + C.$

50. $x\tan x + \ln|\cos x| + C$. **51.** $-\dfrac{x^2}{2} + x\tan x + \ln|\cos x| + C$. **52.** $\dfrac{x}{2}(\cos\ln x + \sin\ln x) + C$. **53.** $\dfrac{x}{2}(\sin\ln x - \cos\ln x) + C$. **54.** $\ln|5 + 2x| + C$. **55.** $-\dfrac{1}{3}\ln|2 - 3x| + C$.

56. $-\dfrac{1}{6(3x+5)^2} + C$. **57.** $\dfrac{1}{8(1-2x)^4} + C$. **58.** $\dfrac{1}{\sqrt{2}}\tan^{-1}\dfrac{x+1}{\sqrt{2}} + C$. **59.** $\ln\sqrt{x^2+2x+5} + \dfrac{1}{2}\tan^{-1}\dfrac{x+1}{2} + C$. **60.** $\ln\sqrt{x^2 - x + 2} + \dfrac{1}{\sqrt{7}}\tan^{-1}\dfrac{2x-1}{\sqrt{7}} + C$. **61.** $\ln\sqrt{x^2+7x+13} - \dfrac{7}{\sqrt{3}}\tan^{-1}\dfrac{2x+7}{\sqrt{3}} + C$. **62.** $\ln|x^2 + 3x - 10| + C$. **63.** $\ln\dfrac{|x+1|}{\sqrt{|2x+1|}} + C$. **64.** $\ln|(x - 1)(x^2 - 4)| + C$. **65.** $\dfrac{x^2}{2} + \ln\left|\dfrac{x-1}{x+1}\right| + C$. **66.** $\dfrac{x^2}{2} + \ln\left|\dfrac{x^2-1}{x}\right| + C$. **67.** $\dfrac{1}{x+3} + \ln|x + 3| + C$. **68.** $\dfrac{1}{3(x+1)^3} - \dfrac{1}{x+1} + C$. **69.** $\dfrac{1}{x} - \dfrac{1}{2x^2} + \ln\left|\dfrac{x}{x+1}\right| + C$.

70. $\dfrac{1}{x-1} + \ln|x + 1| + C$. **71.** $\ln\dfrac{|x|}{\sqrt{x^2+1}} + C$.

72. $\dfrac{1}{4}\ln\left|\dfrac{x-1}{x+1}\right| - \dfrac{1}{2}\tan^{-1}x + C$. **73.** $\dfrac{1}{3}\left(\ln\dfrac{|x-1|}{\sqrt{x^2+x+1}} + \sqrt{3}\tan^{-1}\dfrac{2x+1}{\sqrt{3}}\right) + C$.

74. $\sqrt{\dfrac{1+x}{1-x}} + C$. **75.** $2\sqrt{1+x} + \ln\left|\dfrac{\sqrt{1+x}-1}{\sqrt{1+x}+1}\right| + C$. **76.** $\dfrac{x}{\sqrt{1-x^2}} + C$.

77. $-\dfrac{3}{8}\left(\dfrac{1-x}{1+x}\right)^{3/4} + C$. **78.** $\ln\left|x - 2 + \sqrt{x^2 - 4x}\right| + C$. **79.** $\sin^{-1}\dfrac{x-2}{2} + C$.

80. $\ln\left|x - 2 + \sqrt{x^2-4x-5}\right| + C$. **81.** $\ln(x - 2 + \sqrt{x^2 - 4x + 5}) + C$. **82.** $\sin^{-1}\dfrac{x+2}{3}$.

83. $\sqrt{x^2 - 6x - 1} + 4\ln\left|x - 3 + \sqrt{x^2 - 6x - 1}\right| + C$. **84.** $2\sqrt{1 + 6x - x^2} + 4\sin^{-1}\dfrac{x-3}{\sqrt{10}} + C$. **85.** $-\sqrt{1 + 6x - x^2} + C$. **86.** $x\sqrt{x^2 + 2x + 2} + C$.

87. $x^2\sqrt{1-x^2} + \sin^{-1}x + C$. **88.** $x\sqrt{x^2+2} - \ln(x + \sqrt{x^2 + 2}) + C$. **89.** $-\dfrac{8+4x^2+3x^4}{15} \times \sqrt{1 - x^2} + C$. **90.** $\left(\dfrac{5}{16}x - \dfrac{5}{24}x^3 + \dfrac{1}{6}x^5\right)\sqrt{1 + x^2} - \dfrac{5}{16}\ln(x + \sqrt{x^2 + 1}) + C$.

91. $\ln\left(3 + \tan^2\dfrac{x}{2}\right) + C$. **92.** $\dfrac{2}{3}\tan^{-1}\dfrac{5\tan\frac{x}{2} + 4}{3} + C$. **93.** $\dfrac{1}{5}\ln\left|\dfrac{2\tan\frac{x}{2} - 1}{2\tan\frac{x}{2} + 1}\right| + C$.

94. $\dfrac{2}{\sqrt{15}}\tan^{-1}\dfrac{2\tan\frac{x}{2} + 1}{\sqrt{15}} + C$. **95.** $\ln(1 + \sin^2 x) + C$. **96.** $\dfrac{1}{4}\ln\dfrac{5 - \sin x}{1 - \sin x} + C$.

97. $-x + 2\ln\left|\frac{\tan\frac{x}{2}}{1+\tan\frac{x}{2}}\right| + C.$ **98.** $\sin x - \frac{1}{3}\sin^3 x + C.$ **99.** $-\frac{1}{5}\cos^5 x + \frac{2}{3}\cos^3 x - \cos x + C.$ **100.** $\frac{1}{3}\sin^3 x - \frac{1}{5}\sin^5 x + C.$ **101.** $\frac{1}{7}\cos^7 x - \frac{1}{5}\cos^5 x + C.$ **102.** $\frac{1}{3\cos^3 x} - \frac{1}{\cos x} + C.$ **103.** $\frac{x}{8} - \frac{\sin 4x}{32} + C.$ **104.** $\frac{3}{8}x - \frac{1}{4}\sin 2x + \frac{1}{32}\sin 4x + C.$ **105.** $\frac{x}{16} - \frac{\sin 4x}{64} - \frac{\sin^3 2x}{48} + C.$ **106.** $-\cot x - \frac{1}{3}\cot^3 x + C.$ **107.** $-\frac{1}{3}\cot^3 x + C.$ **108.** $\tan x + \frac{1}{3}\tan^3 x - 2\cot 2x + C.$ **109.** $-\frac{1}{2}\ln\left|\tan\left(x - \frac{\pi}{4}\right)\right| + C.$ **110.** $-\frac{\cos 6x}{12} - \frac{\cos 4x}{8} + C.$ **111.** $-\frac{\cos 6x}{12} + \frac{\cos 4x}{8} + C.$ **112.** $\frac{\sin 10x}{20} + \frac{\sin 4x}{8} + C.$ **113.** $-\frac{\sin 25x}{50} + \frac{\sin 5x}{10} + C.$ **114.** $\frac{3}{5}\sin\frac{5}{6}x + 3\sin\frac{x}{6} + C.$ **115.** $\frac{\cos 6x}{24} - \frac{\cos 4x}{16} - \frac{\cos 2x}{8} + C.$

Chapter 10

Integral Calculus. The Definite Integral

10.1 Basic Concepts and Definitions

Area of a curvilinear figure. Suppose we wish to compute the area of the plane figure $aABb$ bounded by the graph of the positive-valued continuous function $f(x)$, the x-axis and the vertical lines $x = a$ and $x = b$ (Fig. 10.1).

Let us divide (or partition) the closed interval $[a, b]$ into n subintervals by choosing points $x_0, x_1, x_2, \dots, x_n$ such that

$$x_0 = a < x_1 < x_2 < \dots < x_{n-1} < x_n = b.$$

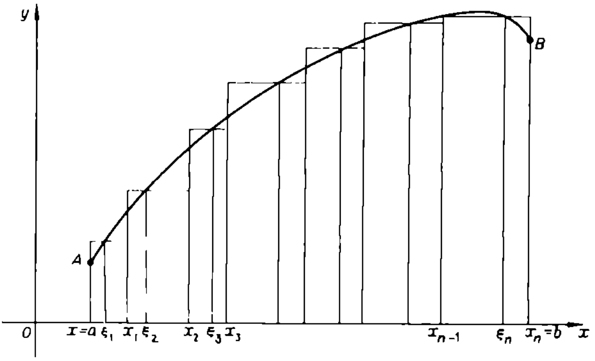


Fig. 10.1

Next we take an arbitrary point ξ_k $(x_{k-1} \leqslant \xi_k \leqslant x_k)$ in each subinterval $[x_{k-1}, x_k]$ and construct n rectangles with bases $[x_{k-1}, x_k]$ and hights $f(\xi_k)$, $k = 1, 2, \dots, n$. The area of the kth rectangle is

$$\Delta Q_k = f(\xi_k)\Delta x_k,$$

where $\Delta x_k = x_k - x_{k-1}$ is the base length.

By doing so we obtain a polygonal figure made up of n rectangles; its area Q_n is equal to the area of the union of these rectangles so that we can write

$$Q_n = f(\xi_1)\Delta x_1 + f(\xi_2)\Delta x_2 + \ldots + f(\xi_n)\Delta x_n = \sum_{k=1}^{n} f(\xi_k)\Delta x_k.$$

Clearly, if we divide the interval $[a, b]$ into smaller subintervals the number of subintervals increases and the lengths of these subintervals become smaller so that the polygonal figure thus obtained comes closer to the plane figure *aABb*.

Let $\lambda = \max\limits_{1 \leqslant k \leqslant n} \Delta x_k$ be the largest of the lengths of the subintervals $[x_{k-1}, x_k]$, $k = 1, 2, \ldots, n$. Evidently, the number of subintervals tends to infinity and their lengths Δx_k tend to zero as $\lambda \to 0$ since $0 \leqslant \Delta x_k \leqslant \lambda$ for all $k = 1, 2, \ldots, n$.

Suppose that there exists a finite limit Q of the polygonal figure as $\lambda = \max\limits_{1 \leqslant k \leqslant n} \Delta x_k \to 0$. Then this limit is said to be equal to the area of the plane figure *aABb*, i.e.,

$$Q = \lim_{\lambda \to 0} Q_n = \lim_{\lambda \to 0} \sum_{k=1}^{n} f(\xi_k)\Delta x_k.$$

This limit, if it exists, must be the same for any collection of subintervals $[x_{k-1}, x_k]$ and any choice of points ξ_k in the subintervals. Thus computing the area of the plane curvilinear figure *aABb* becomes equivalent to evaluating a limit of the form

$$\lim_{\substack{\max \Delta x_k \to 0 \\ 1 \leqslant k \leqslant n}} \sum_{k=1}^{n} f(\xi_k)\Delta x_k.$$

Distance travelled by a particle. We wish to find a distance S travelled by a particle during a time interval from $t = t_0$ to $t = T$ provided that the velocity v of the moving particle is described by the function $v = f(t)$.

Let us partition the interval $[t_0, T]$ into n small subintervals by choosing points $t_0, t_1, t_2, \ldots, t_n$ such that

$$t_0 < t_1 < t_2 < \ldots < t_n = T.$$

Suppose that the velocity $v(t)$ remains nearly unchanged on any subinterval $[t_{k-1}, t_k]$ so that it may be thought of as being equal to the value of $v(t)$ at $\tau_k \in [t_{k-1}, t_k]$, i.e., $v = f(\tau_k)$, $k = 1, 2, \ldots, n$. Then the distance travelled by the particle during the time interval $\Delta t_k = [t_k - t_{k-1}]$ is nearly equal to $s_k = f(\tau_k)\,\Delta t_k$; hence we can approximate the distance S_n travelled by the particle during the time interval $[t_0, T]$ as

$$S_n = s_1 + s_2 + \ldots + s_n = f(\tau_1)\,\Delta t_1 + f(\tau_2)\,\Delta t_2$$
$$+ \ldots + f(\tau_n)\,\Delta t_n = \sum_{k=1}^{n} f(\tau_k)\,\Delta t_k.$$

We denote the largest of the subintervals Δt_k, $k = 1, 2, \ldots, n$, by $\lambda = \max\limits_{1 \leqslant k \leqslant n} \Delta t_k$. The number of subintervals increases infinitely and their lengths tend to zero as $\lambda \to 0$. Evaluating the limit of S_n as $\lambda \to 0$, we obtain the exact value S of the distance travelled by the particle from t_0 to T; so we write

$$S = \lim_{\lambda \to 0} \sum_{k=1}^{n} f(\tau_k)\,\Delta t_k.$$

Thus we have arrived at the limits of the same form as before; if these two limits exist they are called the *definite integrals* of the functions $f(x)$ and $f(t)$ denoted by the symbols $\int\limits_a^b f(x)\,dx$ and $\int\limits_{t_0}^T f(t)\,dt$, respectively.

Definition of the definite integral. Let $f(x)$ be a function defined on the closed interval $[a, b]$ $(a < b)$. Proceeding as before, we divide $[a, b]$ into n subintervals by choosing points

$$x_0 = a < x_1 < x_2 < \ldots < x_{n-1} < x_n = b.$$

We call the collection of subintervals of $[a, b]$ the *partition* of $[a, b]$.

Let $\Delta x_k = x_k - x_{k-1} > 0$ be the length of the kth subinterval and ξ_k be a point of the kth subinterval for each $k = 1, 2, \ldots, n$. A collection of points $\xi_1, \xi_2, \ldots, \xi_n$ is called the *selection* for a given partition of $[a, b]$. Given a partition of $[a, b]$ and a selection for this partition we can evaluate the sum

$$S_n = f(\xi_1)\,\Delta x_1 + f(\xi_2)\,\Delta x_2 + \ldots + f(\xi_n)\,\Delta x_n = \sum_{k=1}^{n} f(\xi_k)\,\Delta x_k,$$

where $f(\xi_k)$ is the value of $f(x)$ at the point $\xi_k \in [x_{k-1}, x_k]$. This sum is called the *integral* (*Riemann*) *sum* for $f(x)$ determined by the given partition of $[a, b]$ and the given selection for this partition; hence a value of S_n is dependent on the partition of $[a, b]$ and the selection for the given partition, i.e., it depends on the way of dividing the interval $[a, b]$ into subintervals $[x_{k-1}, x_k]$ and choosing points ξ_k in these subintervals, $k = 1, 2, \ldots, n$.

Let λ be the largest of the lengths of the subintervals $[x_{k-1}, x_k]$, $k = 1, 2, \ldots, n$, i.e., $\lambda = \max\limits_{1 \leqslant k \leqslant n} \Delta x_k$. We say that a number J is the limit of integral sums $\sum\limits_{k=1}^{n} f(\xi_k)\Delta x_k$ for $f(x)$ on $[a, b]$ if given any number $\varepsilon > 0$ there exists a number $\delta > 0$ such that

$$\left| \sum_{k=1}^{n} f(\xi_k)\Delta x_k - J \right| < \varepsilon$$

for every partition of $[a, b]$ with $\Delta x_k < \delta$, $k = 1, 2, \ldots, n$, i.e., for every partition with $\lambda < \delta$, and every selection ξ_k, $k = 1, 2, \ldots, n$.

In symbols, we write

$$J = \lim_{\lambda \to 0} \sum_{k=1}^{n} f(\xi_k)\,\Delta x_k.$$

Notice that the number δ depends on the value of ε; to signify this we shall sometimes write $\delta = \delta(\varepsilon)$.

If given any partition of the closed interval $[a, b]$, $a < b$, and any selection ξ_k, $k = 1, 2, \ldots, n$, the integral sum $\sum_{k=1}^{n} f(\xi_k)\,\Delta x_k$ has the same finite limit J as $\lambda \to 0$, this limit is called the *definite* (or *Riemann*) *integral* of $f(x)$ on $[a, b]$ and denoted by $\int_a^b f(x)\,dx$. Thus by the definition of the definite integral we have

$$J = \int_a^b f(x)\,dx = \lim_{\lambda \to 0} \sum_{k=1}^{n} f(\xi_k)\,\Delta x_k.$$

The numbers a and b are called the *lower limit* and the *upper limit* of the integral, respectively; x is called the *variable of integration.* The function $f(x)$ is called the *integrand* and the expression $f(x)\,dx$ is called the *element of integration.*

It is worth noting that by virtue of the definition the definite integral remains unchanged when at any point c in $[a, b]$ the value $f(x)$ of the integrand is replaced by some other number. In other words, if we replace the integrand $f(x)$ by the function

$$g(x) = \begin{cases} f(x) & \text{for } x \in [a, b],\ x \neq c, \\ A & \text{for } x = c, \end{cases}$$

where $A \neq f(c)$, then

$$\int_a^b f(x)dx = \int_a^b g(x)dx.$$

This is also true if $f(x)$ is modified at any finite number of points in $[a, b]$.

The definition above applies only if $a < b$; it is also convenient to include the cases $a = b$ and $b < a$. We put $\int_a^a f(x)\,dx = 0$ for $b = a$ and

$$\int_b^a f(x)\,dx = -\int_a^b f(x)\,dx \text{ for } b < a.$$

Example. Compute $\int_a^b dx$.

◄ By the definition we have

$$\int_a^b dx = \lim_{\lambda \to 0} \sum_{k=1}^{n} \Delta x_k = \lim_{\lambda \to 0} \sum_{k=1}^{n} (x_k - x_{k-1})$$

$$= \lim_{\lambda \to 0} [(x_1 - x_0) + (x_2 - x_1) + \ldots + (x_{n-1} - x_{n-2})$$

$$+ (x_n - x_{n-1})] = \lim_{\lambda \to 0} (x_n - x_0) = \lim_{\lambda \to 0} (b - a) = b - a. \quad ►$$

Integrable functions. A function $f(x)$ defined on the closed interval $[a, b]$ is said to be *integrable* on $[a, b]$ if there exists the definite integral $\int_a^b f(x)\,dx$ on $[a, b]$.

Theorem 10.1. *If the function $f(x)$ is integrable on the closed interval $[a, b]$ then $f(x)$ is bounded on $[a, b]$.*

◄ Let $f(x)$ be unbounded on $[a, b]$. We divide $[a, b]$ into n subintervals $[x_{k-1}, x_k]$, $k = 1, 2, \ldots, n$. Since $f(x)$ is unbounded on $[a, b]$ there exists a subinterval, say $[x_0, x_1]$, where $f(x)$ is unbounded. Consider the selection $\xi_1, \xi_2, \ldots, \xi_n$ and the integral sum

$$S_n = \sum_{k=1}^{n} f(\xi_k)\,\Delta x_k = f(\xi_1)\,\Delta x_1 + \sum_{k=2}^{n} f(\xi_k)\,\Delta x_k.$$

Clearly, by a suitable choice of ξ_1 in $[x_0, x_1]$ where $f(x)$ is unbounded it is possible to make the modulus $|S_n|$ of S_n arbitrarily large provided that the points $\xi_2, \xi_3, \ldots, \xi_n$ are the same for all selections, i.e., provided that the sum $\sum_{k=2}^{n} f(\xi_k)\,\Delta x_k$ remains the same for the chosen ξ_1 in $[x_0, x_1]$. This implies that the integral sum S_n has no finite limit as $\max_{1 \leqslant k \leqslant n} \Delta x_k \to 0$, i.e., $f(x)$ is not integrable on $[a, b]$. Whence it follows that if $f(x)$ is integrable on $[a, b]$ then $f(x)$ is bounded on $[a, b]$. ►

Remark. If a function is bounded on $[a, b]$ it is not necessarily integrable on this interval; in other words, a function can be bounded on $[a, b]$ but not integrable on $[a, b]$. For example, the *Dirichlet's function*

$$f(x) = D(x) = \begin{cases} 1 & \text{for rational } x \\ 0 & \text{for irrational } x \end{cases}$$

is bounded on the closed interval $[0, 1]$ since $|f(x)| \leqslant 1$ for all x in $[0, 1]$; however $f(x)$ is not integrable on $[0, 1]$.

◄ Indeed, the integral sum $S_n = \sum_{k=1}^{n} f(\xi_k)\,\Delta x_k$ becomes $S_n = \sum_{k=1}^{n} 1 \cdot \Delta x_k = \sum_{k=1}^{n} \Delta x_k = 1$ for every selection of rational points ξ_k, $k = 1$, 2, ..., n, and $S_n = \sum_{k=1}^{n} 0 \cdot \Delta x_k = 0$ for every selection of irrational points ξ_k. Hence, given any arbitrarily small $\lambda = \max_{1 \leqslant k \leqslant n} \Delta x_k$ the integral sum S_n is equal either to 1 or to 0; this means that S_n has no limit as $\lambda \to 0$, i.e., $f(x)$ is not integrable on [0, 1]. ►

We shall give without proof three theorems that outline sufficient conditions for a function to be integrable on a closed interval.

Theorem 10.2. *If the function $f(x)$ is continuous on the closed interval $[a, b]$ then $f(x)$ is integrable on $[a, b]$.*

For example, the function $f(x) = e^{-x^2}$ is continuous on $[0, a]$ where a is an arbitrary number and, consequently, $f(x)$ is integrable on $[0, a]$, i.e., there exists the definite integral $\int_a^{+\infty} e^{-x^2}\,dx$ of this function.

Theorem 10.3. *If the function $f(x)$ is defined and monotone on the closed interval $[a, b]$ then $f(x)$ is integrable on $[a, b]$.*

It is worth noting that all values of the function $f(x)$ which is monotone on $[a, b]$ lie between the numbers $f(a)$ and $f(b)$; so $f(x)$ is bounded on $[a, b]$.

Theorem 10.4. *Let a function $f(x)$ be bounded on a closed interval $[a, b]$ and let $f(x)$ have a finite number of discontinuities (of the first or second kind) on $[a, b]$. Then $f(x)$ is integrable on $[a, b]$.*

For example, the function

$$f(x) = \begin{cases} \sin\dfrac{1}{x} & \text{for } x \neq 0 \\ 1 & \text{for } x = 0 \end{cases}$$

is integrable on the closed interval [0, 1] since $|f(x)| \leqslant 1$ for all x in [0, 1], i.e., $f(x)$ is bounded on [0, 1], and $f(x)$ has the only discontinuity (of the second kind) at $x = 0$.

10.2 Properties of the Definite Integral

We shall derive some properties of definite integrals assuming that all functions in question are continuous and, consequently, integrable on a closed interval $[a, b]$.

(1) The definite integral is dependent only on its lower and upper limits, i.e., on the numbers a and b, and on the integrand $f(x)$; however, it is independent of the variable of integration. So the definite integral remains unchanged when x is replaced by any other symbol, i.e.,

$$\int_a^b f(x)\,dx = \int_a^b f(t)\,dt = \int_a^b f(u)\,du.$$

(2) The constant factor can be taken outside the integral sign so that

$$\int_a^b A f(x)\,dx = A \int_a^b f(x)\,dx, \quad A = \text{const.}$$

◄ By the definition of the definite integral we have

$$\int_a^b A f(x)\,dx = \lim_{\lambda\to 0} \sum_{k=1}^{n} A f(\xi_k)\,\Delta x_k$$

$$= A \lim_{\lambda\to 0} \sum_{k=1}^{n} f(\xi_k)\,\Delta x_k = A \int_a^b f(x)\,dx. \quad ►$$

(3) The definite integral of the sum (difference) of two functions is equal to the sum (difference) of the integrals of these functions, i.e.,

$$\int_a^b [f_1(x) \pm f_2(x)]\,dx = \int_a^b f_1(x)\,dx \pm \int_a^b f_2(x)\,dx.$$

◄

$$\int_a^b [f_1(x) \pm f_2(x)]\,dx = \lim_{\lambda\to 0} \sum_{k=1}^{n} [f_1(\xi_k) \pm f_2(\xi_k)]\Delta x_k$$

$$= \lim_{\lambda\to 0} \left[\sum_{k=1}^{n} f_1(\xi_k)\,\Delta x_k \pm \sum_{k=1}^{n} f_2(\xi_k)\,\Delta x_k \right]$$

$$= \lim_{\lambda\to 0} \sum_{k=1}^{n} f_1(\xi_k)\,\Delta x_k \pm \lim_{\lambda\to 0} \sum_{k=1}^{n} f_2(\xi_k)\,\Delta x_k$$

$$= \int_a^b f_1(x)\,dx \pm \int_a^b f_2(x)\,dx. \quad ►$$

Corollary. The *linearity property* of the definite integral:

$$\int_a^b [A_1 f_1(x) + A_2 f_2(x)]\,dx = A_1 \int_a^b f_1(x)\,dx + A_2 \int_a^b f_2(x)\,dx,$$

where A_1 and A_2 are arbitrary constants.

(4) Given any numbers a, b and c there holds

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$$

provided that both integrals on the right exist.

This property is often called the *interval union property.*

◄ We distinguish two cases.

(a) Let $a < c < b$. Then by the definition of the definite integral we have

$$\int_a^b f(x)\,dx = \lim_{\lambda\to 0} \sum_{k=1}^{n} f(\xi_k)\,\Delta x_k.$$

Since the definite integral is independent of the partition of $[a, b]$ we can include the point c into the selection for the given partition of $[a, b]$ by choosing the partition (Fig. 10.2) as

$$x_0 = a < x_1 < x_2 < \dots < x_m = c < x_{m+1} < \dots < x_n = b.$$

Then the integral sum $\sum_{k=1}^{n} f(\xi_k)\,\Delta x_k$ associated with this partition of $[a, b]$ can be split into two integral sums as

$$\sum_{k=1}^{n} f(\xi_k)\,\Delta x_k = \sum_{k=1}^{m} f(\xi_k)\,\Delta x_k + \sum_{k=m+1}^{n} f(\xi_k)\,\Delta x_k,$$

where the sums on the right are associated with the partitions of $[a, c]$ and of $[c, b]$, respectively.

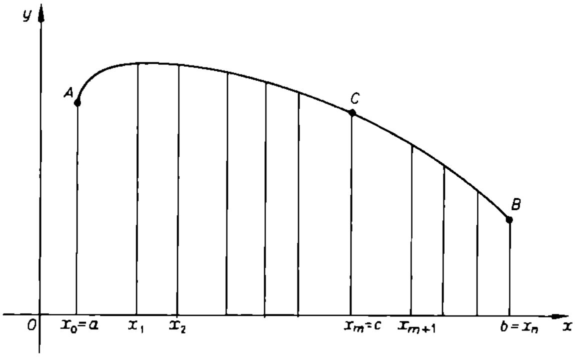


Fig. 10.2

Evaluating the limit as $\lambda = \max\limits_{1 \leqslant k \leqslant n} \Delta x_k \to 0$, we obtain

$$\int_a^b f(x)\,dx = \lim_{\lambda \to 0} \sum_{k=1}^{n} f(\xi_k)\,\Delta x_k = \lim_{\lambda \to 0} \sum_{k=1}^{m} f(\xi_k)\,\Delta x_k$$

$$+ \lim_{\lambda \to 0} \sum_{k=m+1}^{n} f(\xi_k)\,\Delta x_k = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

(b) Now let $a < b < c$. As before we have

$$\int_a^c f(x)\,dx = \int_a^b f(x)\,dx + \int_b^c f(x)\,dx.$$

Whence

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx - \int_b^c f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

Given $f(x) > 0$ and $a < c < b$ this property implies that the area of the curvilinear figure $aABb$ is equal to the sum of the areas of the curvilinear figures $acCA$ and $cbBC$ as easily seen from Fig. 10.2 ►

(5) Let the functions $f(x)$ and $g(x)$ be such that $f(x) \leqslant g(x)$ on $[a, b]$. Then

$$\int_a^b f(x)\,dx \leqslant \int_a^b g(x)\,dx.$$

This means that integration preserves inequalities between functions.
◄ Since $f(x) \leqslant g(x)$ at every point x in $[a, b]$ then given any partition of $[a, b]$ and any selection ξ_k, $k = 1, 2, \ldots, n$, there holds

$$\sum_{k=1}^{n} f(\xi_k)\,\Delta x_k \leqslant \sum_{k=1}^{n} g(\xi_k)\,\Delta x_k.$$

Evaluating the limit as $\lambda = \max\limits_{1\leqslant k\leqslant n} \Delta x_k \to 0$, we get for $a \leqslant b$

$$\int_a^b f(x)\,dx \leqslant \int_a^b g(x)\,dx. \quad ►$$

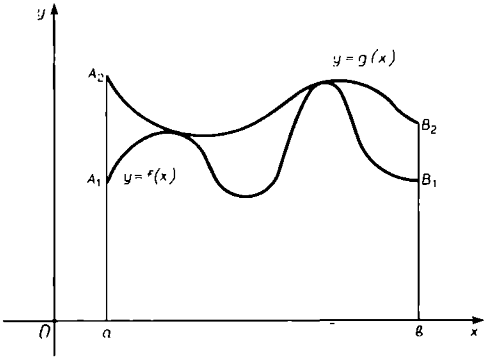


Fig. 10.3

Remark. If $f(x) \geqslant 0$ and $g(x) \geqslant 0$ on $[a, b]$ this property means that the area of the curvilinear figure abB_1A_1 does not exceed the area of abB_2A_2 (Fig. 10.3). In particular, this property implies that

$$\int_a^b f(x)\,dx \geqslant 0 \quad \left(\int_a^b f(x)\,dx \leqslant 0\right)$$

for $f(x) \geqslant 0$ ($f(x) \leqslant 0$) on $[a, b]$ where $a < b$.

(6) Let $a < b$. Then

$$\left| \int_a^b f(x)\,dx \right| \leqslant \int_a^b |f(x)|\,dx.$$

◄ Evaluating the integral of

$$-|f(x)| \leqslant f(x) \leqslant |f(x)|$$

from a to b, we obtain

$$-\int_a^b |f(x)|\,dx \leqslant \int_a^b f(x)\,dx \leqslant \int_a^b |f(x)|\,dx$$

so that

$$\left| \int_a^b f(x)\,dx \right| \leqslant \int_a^b |f(x)|\,dx. \quad ►$$

(7) Let m and M be the minimum and maximum values of $f(x)$, respectively, on $[a, b]$ where $a < b$. Then

$$m(b - a) \leqslant \int_a^b f(x)\,dx \leqslant M(b - a).$$

◄ Since $m \leqslant f(x) \leqslant M$ for all x in $[a, b]$ Property 5 yields

$$\int_a^b m\,dx \leqslant \int_a^b f(x)\,dx \leqslant \int_a^b M\,dx.$$

Observe that

$$\int_a^b m\,dx = m \int_a^b dx = m(b - a)$$

and

$$\int_a^b M\,dx = M \int_a^b dx = M(b - a).$$

Whence

$$m(b - a) \geqslant \int_a^b f(x)dx \leqslant M(b - a). \quad ►$$

Remark. Given a function $f(x) > 0$ on $[a, b]$ where $a < b$, this property implies that the area Q of curvilinear figure $abBA$ lies between the areas Q_1 and Q_2 of the rectangles abB_1A_1 and abB_2A_2 (Fig. 10.4), i.e., $Q_1 \leqslant Q \leqslant Q_2$.

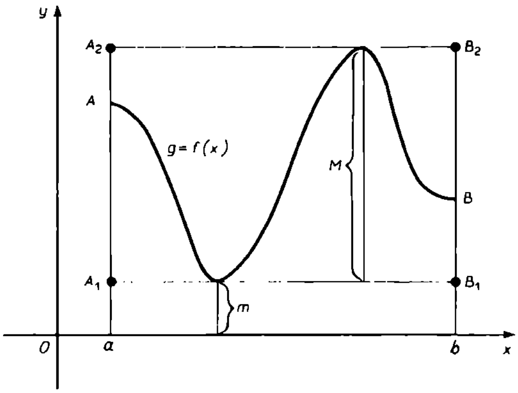


Fig. 10.4

Examples. (1) Evaluate the integral $\int\limits_0^{2\pi} \frac{dx}{\sqrt{10+6\sin x}}$.

◄ Since

$$m = \min_{0\leqslant x\leqslant 2\pi} \frac{1}{\sqrt{10+6\sin x}} = \frac{1}{\sqrt{10+6\sin x}}\bigg|_{x=\frac{\pi}{2}} = 0.25$$

and

$$M = \max_{0\leqslant x\leqslant 2\pi} \frac{1}{\sqrt{10+6\sin x}} = \frac{1}{\sqrt{10+6\sin x}}\bigg|_{x=\frac{3}{2}\pi} = 0.50.$$

Then Property 7 yields

$$2\pi \times 0.25 \leqslant \int\limits_0^{2\pi} \frac{dx}{\sqrt{10+6\sin x}} \leqslant 2\pi \times 0.50$$

and

$$\frac{\pi}{2} \leqslant \int\limits_0^{2\pi} \frac{dx}{\sqrt{10+6\sin x}} \leqslant \pi.$$ ►

(2) Evaluate which of the integrals

$$\int\limits_0^1 e^{-x^2}dx \quad \text{and} \quad \int\limits_0^1 e^{-x}dx$$

is larger without direct computations of their values.

◄ We have $x^2 \leqslant x$ for all x in $[0, 1]$; whence $-x \leqslant -x^2$. Since $e > 1$ then $e^{-x} \leqslant e^{-x^2}$ and Property 5 yields

$$\int_0^1 e^{-x^2}\,dx \geqslant \int_0^1 e^{-x}dx. \quad ►$$

10.3 Fundamental Theorems for Definite Integrals

Mean value theorem. We start with a theorem which states that every continuous function on a closed interval attains a specific value called the *mean value of the function.*

Theorem 10.5 (mean value theorem). *Let $f(x)$ be a continuous function on a closed interval $[a, b]$. Then there exists at least one point ξ in $[a, b]$ such that*

$$\int_a^b f(x)\,dx = (b - a)f(\xi), \quad a \leqslant \xi \leqslant b.$$

◄ Since $f(x)$ is continuous on $[a, b]$ it has the minimum value m and the maximum value M on $[a, b]$ (see Fig. 10.4). Then Property 7 yields

$$m(b - a) \leqslant \int_a^b f(x)\,dx \leqslant M(b - a).$$

Whence

$$m \leqslant \frac{\int_a^b f(x)\,dx}{b - a} \leqslant M$$

because $b - a > 0$.

Put

$$\frac{1}{b - a}\int_a^b f(x)\,dx = \mu,$$

where $m \leqslant \mu \leqslant M$.

Since $f(x)$ is continuous on $[a, b]$, it attains all intermediate values between m and M. This implies that there exists a point $x = \xi$ in $[a, b]$ such that $f(\xi) = \mu$, i.e.,

$$\frac{\int_a^b f(x)\,dx}{b - a} = f(x) \text{ or } \int_a^b f(x)\,dx = (b - a)f(\xi), \quad a \leqslant \xi \leqslant b. \quad ►$$

Remark. For $a < b$ we have

$$a \leqslant \xi \leqslant b \Leftrightarrow 0 \leqslant \xi - a \leqslant b - a \Leftrightarrow 0 \leqslant \frac{\xi - a}{b - a} \leqslant 1.$$

Put

$$\frac{\xi - a}{b - a} = \theta, \quad 0 \leqslant \theta \leqslant 1.$$

Then $\xi = a + (b - a)\theta$ and we rewrite the conclusion of this theorem as

$$\int_a^b f(x)\,dx = (b - a)f[a + (b - a)\theta], \quad 0 \leqslant \theta \leqslant 1.$$

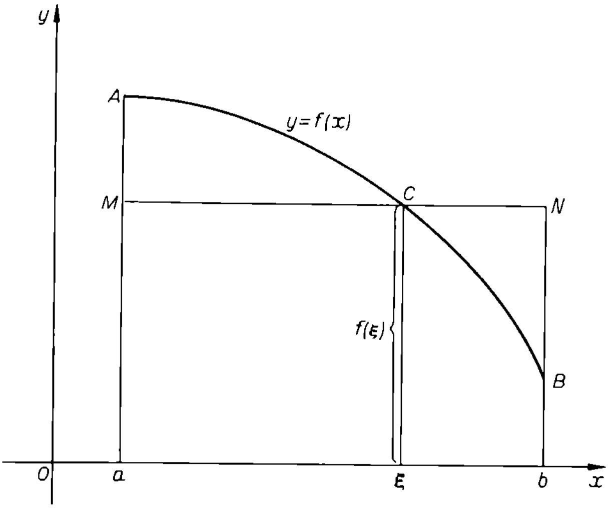


Fig. 10.5

To interpret the mean value theorem for integrals we look at Fig. 10.5. The curvilinear figure *abBA* is bounded by the graph *AB* of the function $f(x)$ which is nonnegative on $[a, b]$ $(a < b)$, the x-axis and the vertical lines $x = a$ and $x = b$. We let Q_1 denote the area of *abBA;* then $\int_a^b f(x)\,dx = Q_1$. For the rectangle *abNM* the base is $[a, b]$ and the height is equal to the ordinate of the point $C(\xi, f(\xi))$ so the area Q_2 of *abNM* is $Q_2 = (b - a)f(\xi)$. The mean value theorem tells us that there exists a point $C(\xi, f(\xi))$ on the graph *AB* such that $Q_1 = Q_2$.

The number

$$M[f(x)] = \frac{1}{b-a}\int_a^b f(x)dx$$

is called the *mean value* of the function $f(x)$ on $[a, b]$.

If a function $f(x)$ is continuous on $[a, b]$ then there exists a point ξ in $[a, b]$ such that $M[f(x)] = f(\xi)$.

Example. Compute the mean value of $f(x) = \sin x$ on $[0, \pi]$.

◄ By virtue of the above definition we have

$$M[\sin x] = \frac{1}{\pi - 0}\int_0^\pi \sin x\,dx = \frac{1}{\pi}(-\cos\pi + \cos 0) = \frac{2}{\pi}. \quad ►$$

We have applied here the *Newton-Leibniz theorem* which will be derived in this section.

Fundamental theorems of calculus. Let $f(x)$ be a continuous function on $[a, b]$. We choose an arbitrary point in $[a, b]$ and consider the definite integral $\int_a^b f(x)\,dx$. Since the definite integral is independent of x we can substitute t for x and write

$$\int_a^x f(x)\,dx = \int_a^x f(t)\,dt.$$

Since $f(x)$ is continuous on $[a, b]$, this integral exists at every x in $[a, b]$. Consequently, the integral becomes a function of its upper limit x. We let $F(x)$ denote this function, i.e.,

$$F(x) = \int_a^x f(t)\,dt.$$

Theorem 10.6 (first fundamental theorem of calculus). *Let $f(x)$ be continuous on $[a, b]$. Then the function $F(x) = \int_a^x f(t)dt$ has a derivative at every point x in $[a, b]$ and $F'(x) = f(x)$.*

In other words, the derivative of the definite integral with respect to its upper limit is equal to the value of the integrand at the upper limit of the integral.

◄ Consider the increment $\Delta x \neq 0$ such that $x + \Delta x \in [a, b]$. This increment gives the increment ΔF in $F(x)$. By virtue of the interval union property (Property 4) of the definite integral we have

$$\Delta F = F(x + \Delta x) - F(x) = \int_a^{x+\Delta x} f(t)dt - \int_a^x f(t)dt = \int_a^{x+\Delta x} f(t)\,dt$$

$$+ \int_x^a f(t)\,dt = \int_x^a f(t)\,dt + \int_a^{x+\Delta x} f(t)\,dt = \int_x^{x+\Delta x} f(t)\,dt.$$

Applying the mean value theorem, we obtain

$$\Delta F = (x + \Delta x - x)f(x + \theta\Delta x) = \Delta x f(x + \theta\,\Delta x).$$

Whence

$$\frac{\Delta F}{\Delta x} = f(x + \theta\,\Delta x),\ 0 \leqslant \theta \leqslant 1.$$

Observe that $f(x)$ is continuous at every point x in $[a, b]$. Then evaluating the limit of the above ratio as $\Delta x \to 0$, we get

$$\lim_{\Delta x \to 0} \frac{\Delta F}{\Delta x} = \lim_{\Delta x \to 0} f(x + \theta\,\Delta x) = f(x),$$

that is,

$$F'(x) = f(x) \text{ or } \left(\int_a^x f(t)\,dt\right)' = f(x) \text{ for all } x \text{ in } [a, b]. \quad ►$$

Remark. If $f(x)$ is a continuous function on $[a, b]$ then for every x in $[a, b]$ there holds

$$\left(\int_x^b f(t)\,dt\right)' = \left(-\int_b^x f(t)\,dt\right)' = -\left(\int_b^x f(t)\,dt\right)' = -f(x).$$

For example, $\left(\int_0^x e^{-t^2}dt\right)' = +e^{-x^2}$ and $\left(\int_x^0 e^{-t^2}dt\right)' = -e^{-x^2}$.

Theorem 10.7 (second fundamental theorem of calculus). *Let $f(x)$ be a continuous function on a closed interval $[a, b]$. Then $f(x)$ has an antiderivative on $[a, b]$ and, consequently, $f(x)$ has the indefinite integral.*

◄ Since $f(x)$ is continuous on $[a, b]$ there exists the definite integral $\int_a^x f(t)\,dt$ for every x in $[a, b]$, i.e., there exists the function $F(x) = \int_a^x f(t)\,dt$ such that $F'(x) = f(x)$ for all x in $[a, b]$. This means that $F(x)$ is an antiderivative of $f(x)$ on $[a, b]$. Whence it follows that the indefinite integral of the function $f(x)$ continuous on $[a, b]$ admits a representation of the form

$$\int f(x)\,dx = \int_a^x f(t)\,dt + C,$$

where C is an arbitrary constant. ►

Newton-Leibniz theorem. We now turn our attention to a theorem which is helpful in computations of definite integrals.

Theorem 10.8 (Newton-Leibniz theorem). *Let $f(x)$ be a continuous function on a closed interval $[a, b]$ and let $F(x)$ be an antiderivative of $f(x)$ on $[a, b]$. Then*

$$\int_a^b f(x)\,dx = F(b) - F(a).$$

◄ Consider the function

$$\Phi(x) = \int_a^x f(t)\,dt, \quad x \in [a, b].$$

This function is an antiderivative of $f(x)$ on $[a, b]$. Recall that any two antiderivatives of a given function differ from each other by a constant; hence there exists a constant C such that

$$\Phi(x) = F(x) + C \quad \text{or} \quad \int_a^x f(t)\,dt = F(x) + C$$

for all x in $[a, b]$.

Put $x = a$. Then $\int_a^a f(t)\,dt = F(a) + C$. Since $\int_a^a f(t)\,dt = 0$, we have $F(a) + C = 0$; whence $C = -F(a)$. Thus we arrive at

$$\int_a^x f(t)\,dt = F(x) - F(a).$$

Put now $x = b$. Then $\int_a^b f(t)\,dt = F(b) - F(a)$ and, substituting x for t, we arrive at the desired formula

$$\int_a^b f(x)\,dx = F(b) - F(a). \quad ►$$

Remark. Using the notation $F(b) - F(a) = F(x)\Big|_a^b$, we can write the Newton-Leibniz theorem as

$$\int_a^b f(x)\,dx = F(x)\Big|_a^b,$$

where $f(x) = F'(x)$.

The Newton-Leibniz theorem establishes a close relation between the definite integral of $f(x)$ and its antiderivative $F(x)$ so that the computation of the former becomes the evaluation of the latter.

Examples. Compute the following definite integrals.

(1) $\int_2^4 x\,dx.$

◄ Recall that

$$\int x\,dx = \frac{x^2}{2} + C \text{ and } F(x) = \frac{x^2}{2} + C.$$

Then

$$\int_2^4 x\,dx = \frac{x^2}{2}\Bigg|_2^4 = \frac{4^2}{2} - \frac{2^2}{2} = 6. \quad ►$$

(2) $\int_0^\pi \sin x\,dx.$

◄ We have

$$\int_0^\pi \sin x\,dx = -\cos x\Bigg|_0^\pi = -\cos\pi - (-\cos 0) = 2. \quad ►$$

10.4 Evaluating Definite Integrals

Integration by substitution. Now we consider how a method of integration by substitution applies when we deal with definite integrals.

Theorem 10.9. *Let there be given the integral* $\int_a^b f(x)\,dx$ *where* $f(x)$ *is a continuous function on a closed interval* $[a, b]$ *and let* $x = \varphi(t)$. *Suppose that the function* $\varphi(t)$ *satisfies the following conditions:*

(i) *the function* $\varphi(t)$ *assumes values from* a *to* b *when* t *varies from* α *to* β *so that* $\varphi(\alpha) = a$, $\varphi(\beta) = b$ *and all intermediate values of* $\varphi(t)$ *are in* $[a, b]$;

(ii) *the derivative* $\varphi'(t)$ *of* $\varphi(t)$ *is a continuous function on the closed interval* $[\alpha, \beta]$.

Then

$$\int_a^b f(x)\,dx = \int_\alpha^\beta f[\varphi(t)]\varphi'(t)\,dt.$$

◄ Applying the Newton-Leibniz theorem, we have

$$\int_a^b f(x)\,dx = F(b) - F(a),$$

where $F(x)$ is an antiderivative of $f(x)$ on $[a, b]$, i.e., $F'(x) = f(x)$ for all x in $[a, b]$.

Consider the composite function $\Phi(t) = F[\varphi(t)]$ in t on $[\alpha, \beta]$. By the chain rule of differentiation of a composite function we obtain

$$\Phi'(t) = F'\ [\varphi(t)]\ \varphi'(t) = f\ [\varphi(t)]\varphi'(t).$$

Hence $\Phi(t)$ is an antiderivative of the function $f[\varphi(t)]\ \varphi'(t)$ continuous on $[\alpha, \beta]$. Then the Newton-Leibniz theorem yields the desired result, namely

$$\int_{\alpha}^{\beta} f[\varphi(t)]\ \varphi'(t)\,dt = \Phi(\beta) - \Phi(\alpha)$$

$$= F[\varphi(\beta)] - F[\varphi(\alpha)] = F(b) - F(a) = \int_{a}^{b} f(x)\,dx. \ ►$$

Remark. This method of integration is appropriate only if by a suitable choice of $\varphi(t)$ the integral $\int_{\alpha}^{\beta} f[\varphi(t)]\ \varphi'(t)\,dt$ decomes easier to compute than the original integral $\int_{a}^{b} f(x)\,dx$.

Notice that the method does not require the original variable be substituted for the new one.

We shall illustrate the integration by substitution by means of examples.

Examples. Integrate by substitution the following integrals.

(1) $\int_{0}^{a} \sqrt{a^2 - x^2}\,dx\ (a > 0).$

◄ Put $x = a \sin t$. Then $dx = a \cos t\,dt$ and $\sqrt{a^2 - x^2} = a \sin t$. Substituting $x = 0$ and $x = a$ into $x = a \sin t$, we get two equations $a \sin t = 0$ and $a \sin t = a$; whence $t = 0$ and $t = \pi/2$ are the lower and upper limits, respectively, of the new integral. Finally we have

$$\int_{0}^{a} \sqrt{a^2 - x^2}\,dx = a^2 \int_{0}^{\frac{\pi}{2}} \cos^2 t\,dt = a^2 \int_{0}^{\frac{\pi}{2}} \frac{1 + \cos 2t}{2}\,dt$$

$$= \frac{a^2}{2}\left(t\,\Big|_{0}^{\frac{\pi}{2}} + \frac{1}{2}\sin 2t\,\Big|_{0}^{\frac{\pi}{2}}\right) = \frac{\pi a^2}{4}. \ ►$$

(2) $\int_{1}^{e} \frac{\ln^2 x}{x}\,dx.$

◄ Put $x = e^t$. Since $t = 0$ for $x = 1$, $t = 1$ for $x = e$ and $t = \ln x$,

$$\int_1^e \frac{\ln^2 x}{x}\,dx = \int_0^1 t^2\,dt = \frac{t^3}{3}\Bigg|_0^1 = \frac{1}{3}. \quad ►$$

It is worth noting that sometimes it is convenient to use the substitution $t = \psi(x)$ instead of $x = \varphi(t)$.

(3) $\displaystyle\int_0^{\ln 2} \sqrt{e^x - 1}\,dx.$

◄ Put $t = \sqrt{e^x - 1}$. Then $x = \ln(t^2 + 1)$ and $dt = \dfrac{2t\,dt}{1 + t^2}$. Since $t = 0$ for $x = 0$ and $t = 1$ for $x = \ln 2$, respectively, we have

$$\int_0^{\ln 2} \sqrt{e^x - 1}\,dx = 2\int_0^1 \frac{t^2\,dt}{1 + t^2} = 2\int_0^1 \frac{(1 + t^2) - 1}{1 + t^2}\,dt$$

$$= 2\int_0^1 \left(1 - \frac{1}{1 + t^2}\right)dt = 2\left(t\,\Big|_0^1 - \tan^{-1} t\,\Big|_0^1\right)$$

$$= 2(1 - \tan^{-1} 1) = 2 - \frac{\pi}{2}. \quad ►$$

(4) $\displaystyle\int_0^1 (2x^3 - 1)\sqrt{x^4 - 2x + 1}\,dx.$

◄ Put $t = x^4 - 2x + 1$. Notice that here we do not need to find the function $x = \varphi(t)$. On differentiating $t = x^4 - 2x + 1$, we have $dt = (4x^3 - 2)dx$; whence $(2x^3 - 1)\,dx = \frac{1}{2}\,dt$ and

$$\int_0^1 (2x^3 - 1)\sqrt{x^4 - 2x + 1}\,dx = \frac{1}{2}\int_1^0 \sqrt{t}\,dt = -\frac{1}{3}. \quad ►$$

The following theorem is sometimes useful in simplifying the computations of definite integrals.

Theorem 10.10. *Let $f(x)$ be an integrable function on a closed interval $[-a, a]$ which is symmetric relative to the point 0, where $a > 0$. Then*

$$\int_{-a}^{a} f(x)\,dx = \begin{cases} 2\displaystyle\int_0^a f(x)\,dx & \text{if } f(x) \text{ is even,} \\ 0 & \text{if } f(x) \text{ is odd.} \end{cases}$$

◄ By the interval union property (Property 4) of the definite integral we have

$$\int_{-a}^{a} f(x)\,dx = \int_{-a}^{0} f(x)\,dx + \int_{0}^{a} f(x)\,dx.$$

Put $x = -t$ so that $dx = -dt$ and $t = -x$. Then

$$\int_{-a}^{0} f(x)\,dx = -\int_{a}^{0} f(-t)\,dt = \int_{0}^{a} f(-t)\,dt = \int_{0}^{a} f(-x)\,dx.$$

Hence

$$\int_{-a}^{a} f(x)\,dx = \int_{0}^{a} [f(-x) + f(x)]\,dx.$$

Recall that for an even function $f(x)$ we have $f(x) = f(-x)$ so that $\int_{-a}^{a} f(x)\,dx = 2\int_{0}^{a} f(x)\,dx$. Similarly, for an odd function $f(x)$, i.e., for $f(-x) = -f(x)$, $\int_{-a}^{a} f(x)\,dx = 0$. ►

For example, $\int_{-\pi}^{\pi} \sin^3 x\, e^{\cos x} dx = 0$ since the integrand is an odd function on the closed interval $[-\pi, \pi]$. Indeed,

$$\sin^3(-x)\,e^{\cos(-x)} = -\sin^3 x\, e^{\cos x}, \quad \forall x \in [-\pi, \pi].$$

Integration by parts. We shall extend the method of integration by parts examined in the previous chapter to definite integrals.

Theorem 10.11. *Let the functions $u = u(x)$ and $v = v(x)$ have continuous derivatives $u'(x)$ and $v'(x)$ on a closed interval $[a, b]$. Then*

$$\int_{a}^{b} u\,dv = uv\,\Big|_{a}^{b} - \int_{a}^{b} v\,du.$$

◄ By the hypothesis $uv = u(x)\,v(x)$ is a differentiable function on $[a, b]$. Then $(uv)' = uv' + vu'$, i.e., uv is an antiderivative of the function $uv' + vu'$ on $[a, b]$. Applying the Newton-Leibniz theorem, we obtain

$$\int_{a}^{b} (uv' + vu')\,dx = uv\,\Big|_{a}^{b} \quad \text{and} \quad \int_{a}^{b} uv'\,dx + \int_{a}^{b} vu'\,dx = uv\,\Big|_{a}^{b}.$$

Whence

$$\int_{a}^{b} uv'\,dx = uv\,\Big|_{a}^{b} - \int_{a}^{b} vu'\,dx.$$

By the definition of the differential we have $v'\,dx = dv$ and $u'\,dx = du$. Substituting these relations into the above identity, we obtain the desired result

$$\int_a^b u\,dv = uv\,\Big|_a^b - \int_a^b v\,du. \quad ►$$

Examples. Apply integration by parts to the following integrals.

(1) $\int_0^\pi (\pi - x)\sin x\,dx.$

◄ We have $u\,dv = (\pi - x)\sin x\,dx$. Put $u = \pi - x$; then $dv = \sin x\,dx$, $du = -dx$ and $v = -\cos x$. Integration by parts yields

$$\int_0^\pi (\pi - x)\sin x\,dx = -(\pi - x)\cos x\,\Big|_0^\pi - \int_0^\pi \cos x\,dx$$

$$= (x - \pi)\cos x\,\Big|_0^\pi - \sin x\,\Big|_0^\pi = \pi. \quad ►$$

(2) $\int_1^e \frac{\ln x}{x^2}\,dx.$

◄ Put $u = \ln x$ and $dv = \frac{dx}{x^2}$. Then $du = \frac{dx}{x}$ and $v = -\frac{1}{x}$ and integration by parts gives

$$\int_1^e \frac{\ln x}{x^2}\,dx = -\frac{1}{x}\ln x\,\Big|_1^e + \int_1^e \frac{dx}{x^2} = -\frac{\ln x}{x}\,\Big|_1^e - \frac{1}{x}\,\Big|_1^e$$

$$= -\frac{1}{e} - \frac{1}{e} + 1 = 1 - 2e^{-1}. \quad ►$$

10.5 Computing Areas and Volumes by Integration

Areas of plane figures in Cartesian coordinates. Let $f(x)$ be a continuous nonnegative function on a closed interval $[a, b]$ where $a < b$. Then the area Q of the curvilinear figure $abBA$ (see Fig. 10.6) is

$$Q = \int_a^b f(x)\,dx.$$

For example, if a plane figure is bounded by the parabola $y = x^2$, the vertical line $x = a$ $(a > 0)$ and the x-axis as shown in Fig. 10.7 then its area Q is

$$Q = \int_0^a x^2\,dx = \frac{x^3}{3}\,\Big|_0^a = \frac{a^3}{3}.$$

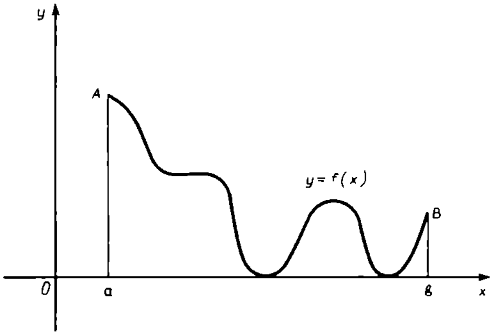


Fig. 10.6

Suppose that $f(x)$ is a negative function on the closed interval $[a, b]$, where $a < b$, i.e., $f(x) < 0$ on $[a, b]$. Then the region (the plane figure) bounded by the graph of $y = f(x)$, the lines $x = a$ and $x = b$ and the x-axis lies below the x-axis (see Fig. 10.8) and $\int_a^b f(x)\,dx < 0$. In this case the area Q of the plane figure $abBA$ is

$$Q = -\int_a^b f(x)\,dx \text{ or } Q = \left|\int_a^b f(x)\,dx\right|.$$

Example. Compute the area of the plane figure bounded by the parabola $y = x^2 - 2x$ and the x-axis (Fig. 10.9).

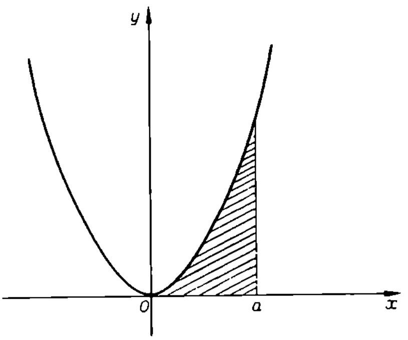


Fig. 10.7

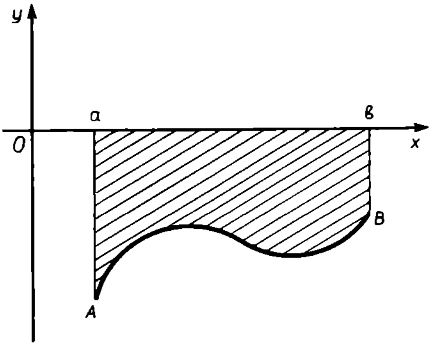


Fig. 10.8

◄ Since $y \leqslant 0$ on $[0, 2]$ the plane figure lies below the x-axis, the desired area is

$$Q = -\int_0^2 (x^2 - 2x)\,dx = \int_0^2 (2x - x^2)\,dx = x^2\Big|_0^2 - \frac{x^3}{3}\Big|_0^2 = \frac{4}{3}$$

and $Q = \frac{4}{3}$. ►

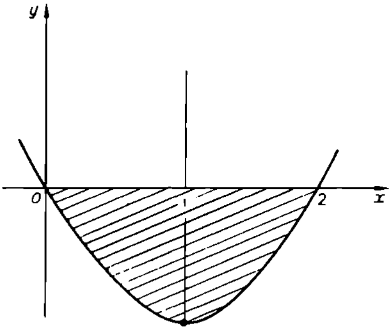


Fig. 10.9

Let a function $f(x)$ change its sign when x passes through a point $c \in (a, b)$. Then the plane figure bounded by the graph of $y = f(x)$, the x-axis and the lines $x = a$ and $x = b$ can be regarded as being made up of two plane figures lying above and below the x-axis (Fig. 10.10). In this case the area Q of the plane figure is

$$Q = Q_1 + Q_2 = \int_a^c f(x)\,dx + \left|\int_c^b f(x)\,dx\right|.$$

For example, for the plane figure bounded by the parabola $y = 1 - x^2$, the line $x = 2$ and the x- and y-axes (Fig. 10.11) is

$$Q = \int_0^1 (1 - x^2)\,dx + \left|\int_1^2 (1 - x^2)\,dx\right|$$

$$= x\Big|_0^1 - \frac{x^3}{3}\Big|_0^1 + \left|x\Big|_1^2 - \frac{x^3}{3}\Big|_1^2\right|$$

$$= 1 - \frac{1}{3} + \left|2 - 1 - \frac{8}{3} + \frac{1}{3}\right| = 2.$$

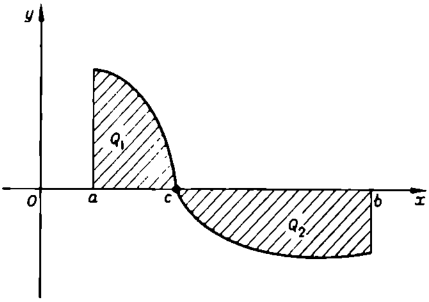


Fig. 10.10

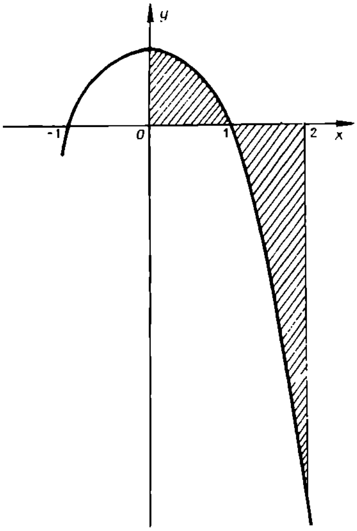


Fig. 10.11

Suppose now that $f(x)$ and $g(x)$ are continuous functions on $[a, b]$, where $a < b$, and $f(x) > g(x) > 0$. Let the graphs of $y = f(x)$ and $y = g(x)$ intersect at the points A and B (Fig. 10.12). Then the area Q of the plane figure bounded by the graphs of $y = f(x)$ and $y = g(x)$ is equal to the difference of the area Q_1 of the plane figure $aACBb$ and the area Q_2 of the plane figure $aADBb$. Thus

$$Q = \int_a^b f(x)\,dx - \int_a^b g(x)\,dx \text{ or } Q = \int_a^b [f(x) - g(x)]\,dx.$$

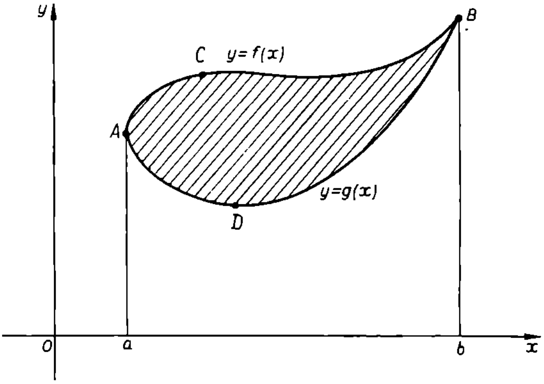


Fig. 10.12

To find the limits a and b of integration one must eliminate y from the system of equations $y + f(x)$ and $y = g(x)$ and solve the equation $f(x) = g(x)$, whose real roots will be the limits sought for.

Example. Compute the area of the plane figure bounded by the parabolas $y = 4x - x^2$ and $y = x^2 - 4x + 6$ (Fig. 10.13).

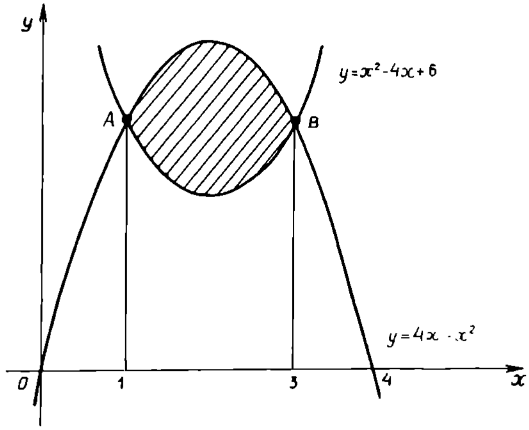


Fig. 10.13

◄ The abscissas of the points A and B where the parabolas intersect are the solutions of the equation $4x - x^2 = x^2 - 4x + 6$ or $x^2 - 4x + 3 = 0$ whose roots are $x_1 = 1$ and $x_2 = 3$; hence the lower and upper limits of integration are $a = 1$ and $b = 3$. Then the desired area Q is

$$Q = \int_1^3 [4x - x^2 - (x^2 - 4x + 6)]\,dx = \int_1^3 (8x - 2x^2 - 6)\,dx$$

$$= 4x^2 \Big|_1^3 - \frac{2}{3}x^3 \Big|_1^3 - 6x \Big|_1^3 = \frac{8}{3}. \quad ►$$

Let a curve AB be given by parametric equations $x = \varphi(t)$ and $y = \psi(t)$ where $\varphi(t)$ and $\psi(t)$ are continuous functions. Suppose that the function $\varphi(t)$ has a continuous derivative $\varphi'(t)$ on the closed interval $[\alpha, \beta]$ and $\varphi(\alpha) = a$ and $\varphi(\beta) = b$ (Fig. 10.14). Evidently, the area Q of the plane figure $abBA$ is

$$Q = \int_a^b y\,dx.$$

If we put $x = \varphi(t)$ and $y = \psi(t)$ then we get the following expression for the area Q of the plane figure specified by parametric equations

$$Q = \int_{\alpha}^{\beta} \psi(t)\, \varphi'(t)\, dt.$$

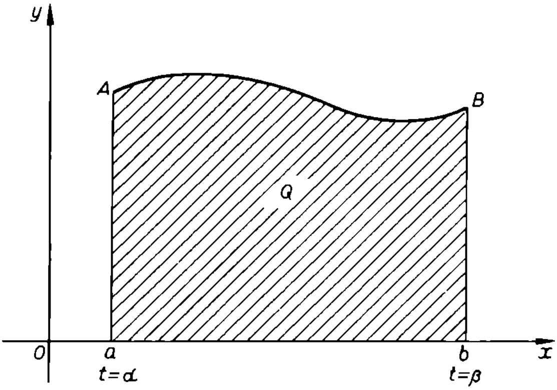


Fig. 10.14

Example. Compute the area of the ellipse given by the parametric equations $x = a \cos t$ and $y = b \sin t$, $0 \leqslant t < 2\pi$ $(a,\ b > 0)$.

◄ Since the ellipse is symmetric relative to the x-axis and to the y-axis it suffices to compute the area of the ellipse in the first quadrant; the desired area is

$$Q = 4 \int_{0}^{a} y\, dx.$$

Put $x = a \cos t$ and $y = b \sin t$ so that $dx = -a \sin t\, dt$. To find the new lower and upper limits of integration we set $x = 0$ and $x = a$ in $x = a \cos t$; this gives $a \cos t = 0$ and $a = a \cos t$. Whence $t_1 = \alpha = \pi/2$ and $t_2 = \beta = 0$. Therefore if x varies from 0 to a t changes from $\pi/2$ to 0 so that

$$Q = 4 \int_{\pi/2}^{0} b \sin t\,(-a \sin t\, dt) = -4ab \int_{\pi/2}^{0} \sin^2 t\, dt$$

$$= 4ab \int_{0}^{\pi/2} \frac{1 - \cos 2t}{2}\, dt = 2ab\left(t \Big|_{0}^{\pi/2} - \frac{1}{2} \sin 2t \Big|_{0}^{\pi/2}\right) = \pi ab. \text{ ►}$$

Sometimes it is convenient to compute areas of plane figures by applying formulas that involve integration with respect to the variable y. In this case the variable x is regarded as a function in y, i.e., $x = g(y)$ where $g(y)$

is a single-valued function continuous on the closed interval $[c, d]$ of the y-axis. The limits c and d of integration with respect to y that are the points of intersection of a given curve with the y-axis can be computed as the roots of the equation $g(y) = 0$ obtained by putting $x = 0$ in the equation $x = g(y)$. Then the area Q of the plane figure bounded by the graph of the curve $x = g(y)$ and the y-axis (Fig. 10.15) is given by the formula

$$Q = \int_c^d g(y)\,dy.$$

Example. Compute the area of the plane figure bounded by the parabola $x = 2 - y - y^2$ and the y-axis (Fig. 10.16).

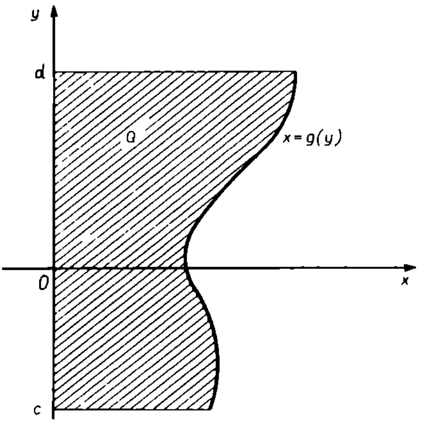


Fig. 10.15

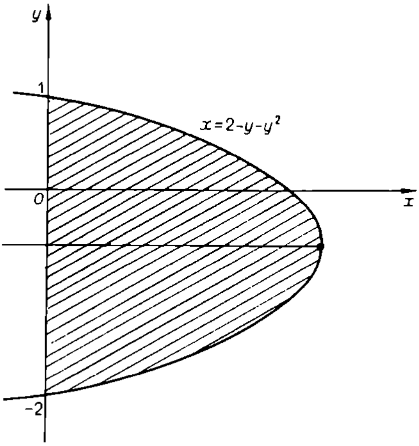


Fig. 10.16

◄ The limits of integration are the points of intersection of the parabola with the y-axis so that putting $x = 0$, we get the equation $2 - y - y^2 = 0$; whence $y_1 = c = -2$ and $y_2 = d = 1$. The desired area Q is

$$Q = \int_{-2}^{1} (2 - y - y^2)\,dy = 2y\Big|_{-2}^{1} - \frac{y^2}{2}\Big|_{-2}^{1} - \frac{y^3}{3}\Big|_{-2}^{1} = 4.5.$$

Problems. (1) Compute the area of the plane figure bounded by the parabola $y^2 = 2x + 1$ and the line $x - y - 1 = 0$;

(2) Compute the area of the plane figure bounded by the curves $y = \sin^{-1} x$ and $y = \cos^{-1} x$ and the x-axis. (*Hint.* Write the equations of the curve in the form $x = g(y)$.)

Area of a plane figure in polar coordinates. Suppose that a curve is specified by its polar equation $\varrho = f(\varphi)$ where $f(\varphi)$ is a continuous nonnegative function on a closed interval $[\alpha, \beta]$. The plane figure bounded by this curve and the two rays that start at the pole and make the angles α and β, respectively, with the polar axis is called the *curvilinear sector* (Fig. 10.17).

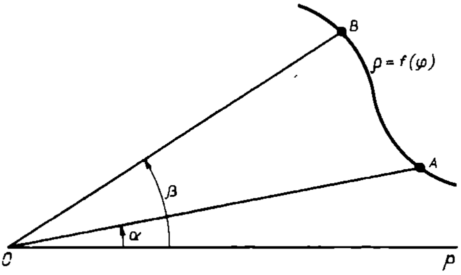


Fig. 10.17

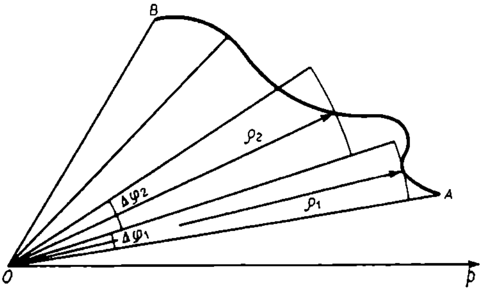


Fig. 10.18

To compute the area of the curvilinear sector $OABO$ we divide the sector into n subsectors by drawing n rays $\varphi = \alpha = \varphi_0$; $\varphi = \varphi_1, \ldots, \varphi_{n-1}$; $\varphi = \beta = \varphi_n$ from the pole. Let $\Delta\varphi_1, \Delta\varphi_2, \ldots, \Delta\varphi_n$ be angles between the rays. We let $\tilde{\varphi}_k$ denote the angle formed by the rays φ_{k-1} and φ_k and $\tilde{\varrho}_k$ the position vector associated with $\tilde{\varphi}_k$. Consider the circular sector of radius $\tilde{\varrho}_k$ with central angle $\Delta\varphi_k$ (Fig. 10.18). The area of this circular sector is equal to $\Delta Q_k = \frac{1}{2}\tilde{\varrho}_k\Delta\varphi_k$ or $\Delta Q_k = \frac{1}{2}f^2(\tilde{\varphi}_k)\Delta\varphi_k$ since $\tilde{\varrho}_k = f(\varphi_k)$.

When we replace each curvilinear subsector by the corresponding circular sector we obtain the plane figure made up of n circular sectors; its area Q_n is

$$Q_n = \sum_{k=1}^{n} \Delta Q_n = \frac{1}{2} \sum_{k=1}^{n} f^2(\tilde{\varphi}_k)\,\Delta\varphi_k.$$

We denote the largest $\Delta\varphi_k$ by $\lambda = \max\limits_{1\leqslant k\leqslant n} \Delta\varphi_k$. If we make n tend to infinity, so that $\lambda \to 0$, i.e., if we divide a given curvilinear sector into smaller subsectors, then the plane figure made up of circular sectors comes closer to the curvilinear sector *OABO*. Thus we may regard the limit of the area Q_n as $\lambda = \max\limits_{1\leqslant k\leqslant n} \Delta\varphi_k \to 0$ as the area of the curvilinear sector *OABO* provided that this limit exists and is independent of a partition of the closed interval $[\alpha, \beta]$ and of a selection $\tilde{\varphi}_k$ associated with a given partition, $k = 1, 2, \ldots, n$, i.e.,

$$Q = \lim_{\lambda\to 0} Q_n = \lim_{\lambda\to 0} \sum_{k=1}^{n} \frac{1}{2} f^2(\tilde{\varphi}_k)\,\Delta\varphi_k.$$

Clearly, the sum $\sum\limits_{k=1}^{n} \frac{1}{2} f^2(\tilde{\varphi}_k)\,\Delta\varphi_k$ is an integral sum for the function $\frac{1}{2} f^2(\varphi)$ which is continuous on $[\alpha, \beta]$ since $f(\varphi)$ is continuous on $[\alpha, \beta]$. Hence, the limit of this sum as $\lambda \to 0$ exists and is equal to the definite integral $\int\limits_{\alpha}^{\beta} \frac{1}{2} f^2(\varphi)\,d\varphi$.

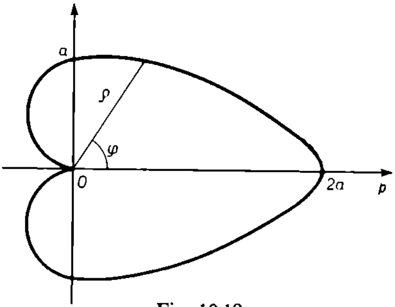


Fig. 10.19

Therefore the area of the curvilinear sector *OABO* is given by

$$Q = \frac{1}{2} \int_{\alpha}^{\beta} f^2(\varphi)\,d\varphi \text{ or } Q = \frac{1}{2} \int_{\alpha}^{\beta} \varrho^2 d\varphi.$$

Example. Compute the area of the plane figure bounded by the *cardioid* $\varrho = a(1 + \cos\varphi)$, $a > 0$ (Fig. 10.19).

◀ The desired area is

$$Q = \frac{a^2}{2}\int_0^{2\pi}(1+\cos\varphi)^2 d\varphi = \frac{a^2}{2}\int_0^{2\pi}(1+2\cos\varphi+\cos^2\varphi)\,d\varphi$$

$$= \frac{a^2}{2}\int_0^{2\pi}\left(1+2\cos\varphi+\frac{1+\cos 2\varphi}{2}\right)d\varphi$$

$$= \frac{a^2}{2}\int_0^{2\pi}\left(\frac{3}{2}+2\cos\varphi+\frac{1}{2}\cos 2\varphi\right)d\varphi = \frac{3}{2}\pi a^2. \quad ▶$$

Volumes of solids. We consider a solid bounded by a closed surface. Let Q be the area of the cross section of the solid by a plane perpendicular to the x-axis (Fig. 10.20). Clearly the value of Q is dependent on the location of the perpendicular plane relative to the x-axis, i.e., this value is a function in x so that $Q = Q(x)$. We assume that $Q(x)$ is a continuous function on a closed interval $[a, b]$. To set up the formula for the volume of the given solid we draw n planes $x = a = x_0$; $x = x_1, x_2, \ldots, x_{n-1}$; $x = b = x_n$ perpendicular to the x-axis. These planes divide the solid into n slices and the closed interval $[a, b]$ into n subintervals $[x_{k-1}, x_k]$, $k = 1, 2, \ldots, n$.

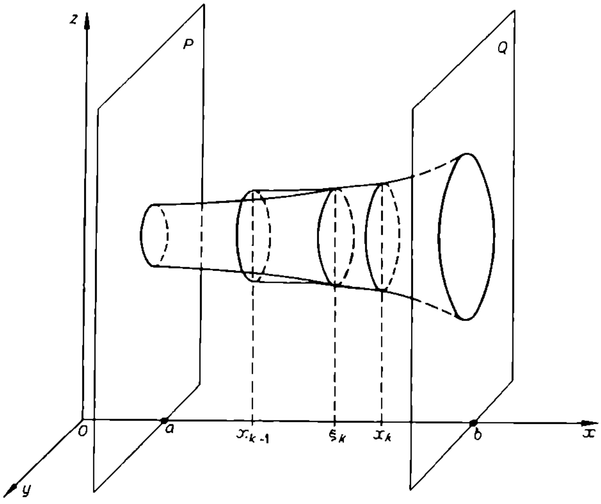


Fig. 10.20

To approximate the volume of the kth slice we select an arbitrary point ξ_k in $[x_{k-1}, x_k]$ and consider the cylinder whose generating line is parallel to the x-axis and directing line is the line of intersection of the solid surface with the perpendicular plane $x = \xi_k$ (see Fig. 10.20). The volume ΔV_k of this cylinder is the product of its height Δx_k and the area $Q(\xi_k)$ of its base so that $\Delta V_k = Q(\xi_k)\,\Delta x_k$, where $\xi_k \in [x_{k-1}, x_k]$. The volume of all n cylinders thus obtained is

$$V_n = \sum_{k=1}^{n} \Delta V_k = \sum_{k=1}^{n} Q(\xi_k)\,\Delta x_k.$$

If the limit of the sum on the right exists as $\lambda = \max\limits_{1 \leqslant k \leqslant n} \Delta x_n \to 0$ we set it equal to the volume of the solid in question so that

$$V = \lim_{\lambda \to 0} \sum_{k=1}^{n} Q(\xi_k)\,\Delta x_k.$$

Notice that this sum is an integral sum for the function $Q(x)$ which is continuous on the closed interval $[a, b]$; hence, the above limit exists and is equal to the definite integral

$$V = \int_a^b Q(x)\,dx. \qquad (*)$$

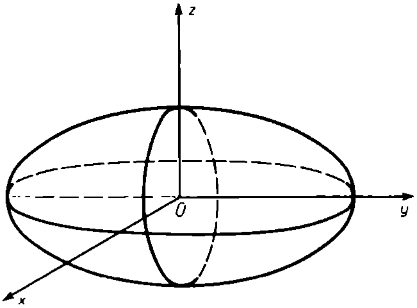


Fig. 10.21

Example. Compute the area of the solid bounded by the ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

◄ The cross section of the ellipsoid with the plane perpendicular to the x-axis at x is the ellipse (Fig. 10.21)

$$\frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 - \frac{x^2}{a^2}$$

or

$$\frac{y^2}{\left(b\sqrt{1-\frac{x^2}{a^2}}\right)^2}+\frac{z^2}{\left(c\sqrt{1-\frac{x^2}{a^2}}\right)^2}=1$$

with the semiaxes

$$b\sqrt{1-\frac{x^2}{a^2}} \text{ and } c\sqrt{1-\frac{x^2}{a^2}}\,.$$

Then the area of the cross section is

$$Q(x)=\pi bc\left(1-\frac{x^2}{a^2}\right).$$

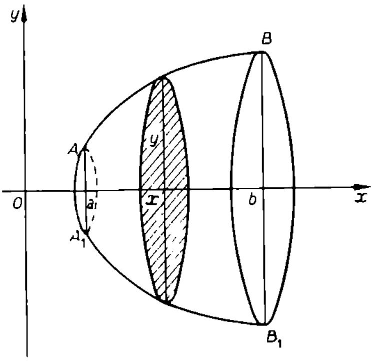


Fig. 10.22

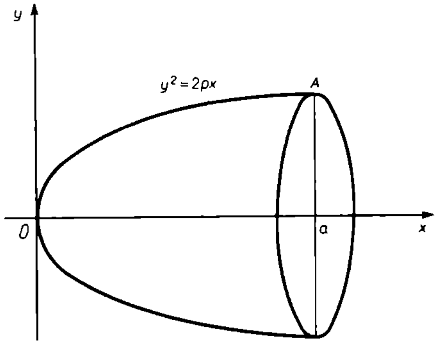


Fig. 10.23

Applying (*), we obtain

$$V=\int_{-a}^{a}\pi bc\left(1-\frac{x^2}{a^2}\right)dx=2\pi bc\int_{0}^{a}\left(1-\frac{x^2}{a^2}\right)dx=\frac{4}{3}\pi abc.$$

When $b=c=a$ the ellipsoid becomes the sphere $x^2+y^2+z^2=a^2$ whose volume is equal to $V=\frac{4}{3}\pi a^3$. ►

Now we consider a solid generated by revolving the curvilinear plane figure *abBA* around the x-axis (Fig. 10.22). The plane figure is bounded by the graph of the curve $y=f(x)$, the lines $x=a$ and $x=b$ where $a<b$ and the x-axis. The solid thus obtained is called the *solid of revolution.*

The cross section of the solid of revolution with the plane perpendicular to the x-axis at x is the circle whose area is $Q(x) = \pi y^2 = \pi f^2(x)$; hence, the volume of the solid of revolution is

$$V = \pi \int_a^b f^2(x)\,dx \text{ or } V = \pi \int_a^b y^2\,dx.$$

Example. Compute the volume of the solid of revolution generated by revolving the arc OA of the parabola $y^2 = 2px$ around the x-axis (Fig. 10.23).

◄ The arc OA is specified by the equation $y = \sqrt{2px}$ where $p > 0$; hence, the desired volume is

$$V = \pi \int_0^a y^2\,dx = \pi \int_0^a 2px\,dx = \pi p a^2. \quad ►$$

10.6 Computing Arc Lengths by Integration

Arc length. We consider the arc $\overset{\smile}{AB}$ with the endpoints A and B (Fig. 10.24). Let $M_1, M_2, \ldots, M_{n-1}$ be points on AB. If we join these points by line segments $AM_1, M_1M_2, \ldots, M_{n-1}B$ with lengths $\Delta s_1, \Delta s_2, \ldots, \Delta s_n$, respectively, we obtain the polygonal line $AM_1M_2 \ldots M_{n-1}B$ inscribed in the arc $\overset{\smile}{AB}$. The length S_n of this polygonal line is

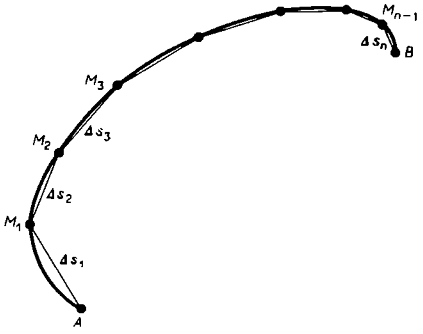


Fig. 10.24

$$S_n = \Delta s_1 + \Delta s_2 + \ldots + \Delta s_n = \sum_{k=1}^{n} \Delta s_k.$$

Definition. The length S of the arc $\overset{\smile}{AB}$ is the limit of the length S_n of the polygonal line inscribed in $\overset{\smile}{AB}$ as the largest of the line segments of this polygonal line tends to zero, i.e.,

$$S = \lim_{\max \Delta s_k \to 0} S_n = \lim_{\max \Delta s_k \to 0} \sum_{k=1}^{n} \Delta s_k$$

provided that this limit exists and is independent of choices of the points $M_1, M_2, \ldots, M_{n-1}$ on the arc $\breve{AB}$.

Arc length of a curve in Cartesian coordinates. Let an arc $\breve{AB}$ be specified by the equation $y = f(x)$ where the function $f(x)$ has a continuous derivative $f'(x)$ on a closed interval $[a, b]$. Consider a partition of $[a, b]$ into n subintervals $[x_{k-1}, x_k]$, $k = 1, 2, \ldots, n$ by choosing points such that

$$x_0 = a < x_1 < x_2 < \ldots < x_{k-1} < x_k < \ldots < x_n = b.$$

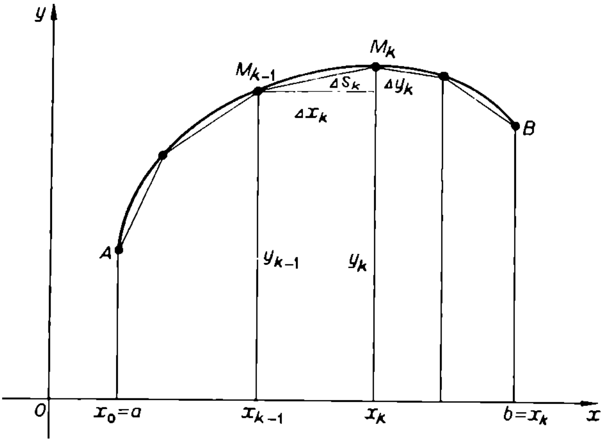


Fig. 10.25

Let $(x_k, f(x_k))$ be a point on $\breve{AB}$ corresponding to the kth subdivision point x_k of the polygonal line. Then the polygonal line inscribed in $\breve{AB}$ is given by the points

$$A = M_0(x_0, f(x_0)),\ M_1(x_1, f(x_1)),\ \ldots,\ B = M_n(x_n, f(x_n)).$$

We let Δs_k denote the length of the kth line segment of this polygonal line and $\Delta x_k = x_k - x_{k-1}$ and $\Delta y_k = f(x_k) - f(x_{k-1})$, $k = 1, 2, \ldots, n$. Then the length of the kth line segment (Fig. 10.25) is

$$\Delta s_k = \sqrt{(\Delta x_k)^2 + (\Delta y_k)^2} = \sqrt{1 + \left(\frac{\Delta y_k}{\Delta x_k}\right)^2}\, \Delta x_k.$$

Applying the mean value theorem for derivatives we obtain

$$\Delta y_k = f(x_k) - f(x_{k-1}) = (x_k - x_{k-1}) \cdot f'(\xi_k) = f'(\xi_k)\, \Delta x_k,$$

where ξ_k is a point in $[x_{k-1}, x_k]$; whence

$$\Delta s_k = \sqrt{1 + [f'(\xi_k)]^2}\, \Delta x_k$$

so that the length of the polygonal line is

$$S_n = \sum_{k=1}^{n} \sqrt{1 + [f'(\xi_k)]^2}\, \Delta x_k. \tag{$*$}$$

Recall that $f'(x)$ is continuous on $[a, b]$. This implies that $\sqrt{1 + [f'(x)]^2}$ is a continuous function on $[a, b]$. Hence the integral sum on the right of (∗) has a limit S as $\max\limits_{1 \leqslant k \leqslant n} \Delta s_k \to 0$ so that

$$S = \lim_{\substack{\max \Delta s_k \to 0 \\ 1 \leqslant k \leqslant n}} \sum_{k=1}^{n} \sqrt{1 + [f'(\xi_k)]^2}\, \Delta x_k = \int_a^b \sqrt{1 + [f'(x)]^2}\, dx$$

or, using the abbreviated notation,

$$S = \int_a^b \sqrt{1 + (y_x')^2}\, dx. \tag{$**$}$$

Example. Compute the arc length of the *catenary* $y = \dfrac{e^x + e^{-x}}{2} = \cosh x$ bounded by the points $A(0, 1)$ and $B(a, \cosh a)$ (Fig. 10.26).

◄ On differentiating the equation of the catenary we have

$$y' = \sinh x.$$

Noting that $\cosh^2 x - \sinh^2 x = 1$, we obtain

$$\sqrt{1 + (y')^2} = \sqrt{1 + \sinh^2 x} = \sqrt{\cosh^2 x} = \cosh x \ (\cosh x > 0)$$

so that

$$S = \int_0^a \cosh x\, dx = \sinh x \Big|_0^a = \sinh a. \ ►$$

Arc length of a curve specified by parametric equations. Let $\overset{\smile}{AB}$ be an arc of a curve specified by the parametric equations

$$x = \varphi(t),\ y = \psi(t),\ t_0 \leqslant t \leqslant T.$$

We also assume that the functions $\varphi(t)$ and $\psi(t)$ have continuous derivatives $\varphi'(t)$ and $\psi'(t)$ on $[t_0, T]$ and $\varphi'(t) \neq 0$ on $[t_0, T]$. The original parametric equations specify the function $y = f(x)$ having a continuous derivative $y_x' = \dfrac{\psi'(t)}{\varphi'(t)}$ on $[t_0, T]$. Then

$$\sqrt{1 + (y_x'^2)}\, dx = \sqrt{[\varphi'(t)]^2 + [\psi'(t)]^2}\, dt$$

and formula (**) yields

$$S = \int_{t_0}^{T} \sqrt{[\varphi'(t)]^2 + [\psi'(t)]^2}\, dt$$

or

$$S = \int_{t_0}^{T} \sqrt{(x_t')^2 + (y_t')^2}\, dt. \qquad (***)$$

Examples. (1) Compute the length of the circumference of the circle with radius R (Fig. 10.27).

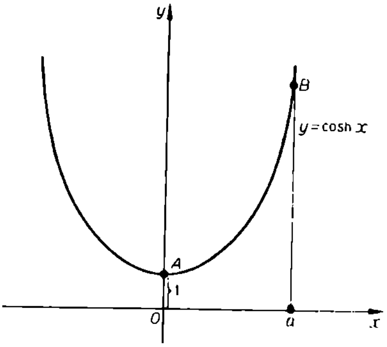


Fig. 10.26

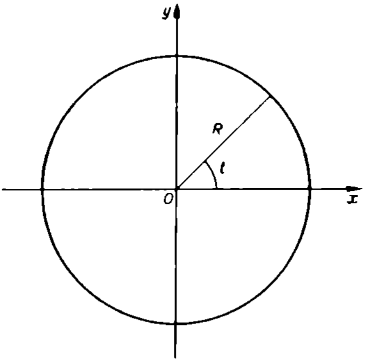


Fig. 10.27

◀ The circle with radius R is given by the parametric equations

$$x = R\cos t,\ y = R\sin t,\ 0 \leqslant t < 2\pi.$$

Then applying formula (***) we obtain

$$S = \int_{0}^{2\pi} \sqrt{(-R\sin t)^2 + (R\cos t)^2}\, dt = R\int_{0}^{2\pi} dt = 2\pi R. \ ▶$$

(2) Find the length of the ellipse given by the parametric equations

$$x = a\cos t,\ y = b\sin t,\ 0 \leqslant t < 2\pi\ (0 < b \leqslant a).$$

◀ Since $x_t' = -a\sin t$ and $y_t' = b\cos t$ and the ellipse is symmetric relative to the coordinate axes, formula (***) gives

$$S = 4\int_{0}^{\pi/2} \sqrt{a^2\sin^2 t + b^2\cos^2 t}\, dt$$

$$= 4\int_{0}^{\pi/2} \sqrt{a^2(1 - \cos^2 t) + b^2\cos^2 t}\, dt$$

$$= 4\int_{0}^{\pi/2} \sqrt{a^2 - (a^2 - b^2)\cos^2 t}\, dt = 4a\int_{0}^{\pi/2} \sqrt{1 - \varepsilon^2\cos^2 t}\, dt,$$

where $\varepsilon = \dfrac{\sqrt{a^2 - b^2}}{a}$ is the eccentricity of the ellipse, $0 \leqslant \varepsilon < 1$.

The definite integral on the right is called the *elliptic integral;* it is worth noting here that the Newton-Leibniz theorem is inapplicable for computing this integral since the antiderivative is not an elementary function. ►

Notice that the substitution $t = \dfrac{\pi}{2} - \tau$ yields

$$\int_0^{\pi/2} \sqrt{1 - \varepsilon^2 \cos^2 t}\, dt = \int_0^{\pi/2} \sqrt{1 - \varepsilon^2 \sin^2 \tau}\, d\tau = \int_0^{\pi/2} \sqrt{1 - \varepsilon^2 \sin^2 t}\, dt.$$

This is the notation mostly used for the elliptic integral.

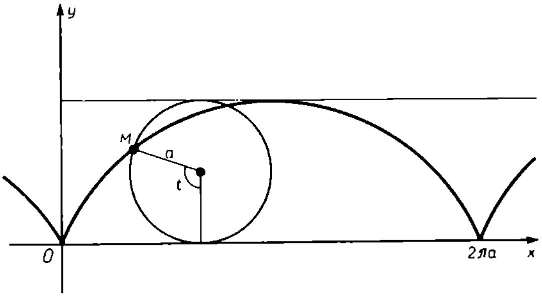


Fig. 10.28

(3) Find the length of one arc of the cycloid specified by the parametric equations (Fig. 10.28)

$$x = a(t - \sin t),\ y = a(1 - \cos t),\ 0 \leqslant t < 2\pi \ (a > 0).$$

◄ Applying formula (***), we have

$$S = a \int_0^{2\pi} \sqrt{(1 - \cos t)^2 + \sin^2 t}\, dt = a \int_0^{2\pi} \sqrt{2 - 2\cos t}\, dt$$

$$= a \int_0^{2\pi} \sqrt{4 \sin^2 \frac{t}{2}}\, dt = 2a \int_0^{2\pi} \left| \sin \frac{t}{2} \right| dt$$

$$= 2a \int_0^{2\pi} \sin \frac{t}{2}\, dt = -4a \cos \frac{t}{2} \Big|_0^{2\pi} = 8a. \text{ ►}$$

Arc length of a curve specified by its polar equation. Let $\breve{AB}$ be an arc of a curve specified by its polar equation $\varrho = f(\varphi)$ on a closed interval $[\alpha, \beta]$ where $f(\varphi)$ has a continuous derivative $f'(\varphi)$. To find the length of the arc $\overset{\smile}{AB}$ we derive the parametric equations of the curve. Recall that the relations between the polar and Cartesian coordinates are given by $x = \varrho\cos\varphi$ and $y = \varrho\sin\varphi$. Replacing ϱ by $f(\varphi)$, we get the parametric equations $x = f(\varphi)\cos\varphi$ and $y = f(\varphi)\sin\varphi$ that describe the given curve. We use here the polar angle φ as the parameter.

On differentiating the parametric equations, we have

$$x'_\varphi = f'(\varphi)\cos\varphi - f(\varphi)\sin\varphi$$

and

$$y'_\varphi = f'(\varphi)\sin\varphi + f(\varphi)\cos\varphi.$$

Squaring both identities and adding $(x'_\varphi)^2$ with $(y'_\varphi)^2$, we get

$$(x'_\varphi)^2 + (y'_\varphi)^2 = [f'(\varphi)]^2 + [f(\varphi)]^2.$$

Then formula (***) gives

$$S = \int_\alpha^\beta \sqrt{[f'(\varphi)]^2 + [f(\varphi)]^2}\, d\varphi$$

or, equivalently,

$$S = \int_\alpha^\beta \sqrt{\varrho^2 + (\varrho')^2}\, d\varphi.$$

Example. Find the length of the cardioid given by the polar equation

$$\varrho = a(1 + \cos\varphi),\ a > 0 \text{ (see Fig. 10.19).}$$

◀ On differentiating the original equation we obtain $\varrho' = -a\sin\varphi = f'(\varphi)$, $0 \leqslant \varphi \leqslant 2\pi$, so that by symmetry

$$S = 2\int_0^\pi \sqrt{a^2(1 + \cos\varphi)^2 + a^2\sin^2\varphi}\, d\varphi$$

$$= 2a\int_0^\pi \sqrt{2(1 + \cos\varphi)}\, d\varphi = 4a\int_0^\pi \sqrt{\cos^2\frac{\varphi}{2}}\, d\varphi$$

$$= 4a\int_0^\pi \cos\frac{\varphi}{2}\, d\varphi = 8a. \quad ▶$$

Now we consider an arc of the curve $y = f(x)$ shown in Fig. 10.29. Suppose that the function $f(x)$ has a continuous derivative $f'(x)$ on the closed interval $[a, b]$. We let $\overset{\smile}{AM}$ denote the arc bounded by the points $A(a, f(a))$

and $M(x, f(x))$. If we think of A as a fixed reference point and of M as a point that can move along the curve the length of the arc $\overset{\smile}{AM}$ becomes a function in x. This length S is expressed as

$$S = \int_a^x \sqrt{1 + [f'(x)]^2}\, dx = \int_a^x \sqrt{1 + [f'(t)]^2}\, dt.$$

Since the integrand on the right is a continuous function on $[a, b]$ we can write

$$\frac{dS(x)}{dx} = \frac{d}{dx}\left(\int_a^x \sqrt{1 + [f'(t)]^2}\, dt\right) = \sqrt{1 + [f'(x)]^2}$$

or

$$\frac{dS}{dx} = \sqrt{1 + \left(\frac{dy}{dx}\right)^2}.$$

Whence the length dS of the arc $\overset{\smile}{AM}$ becomes

$$dS = \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx \text{ or } dS = \sqrt{(dx)^2 + (dy)^2}.$$

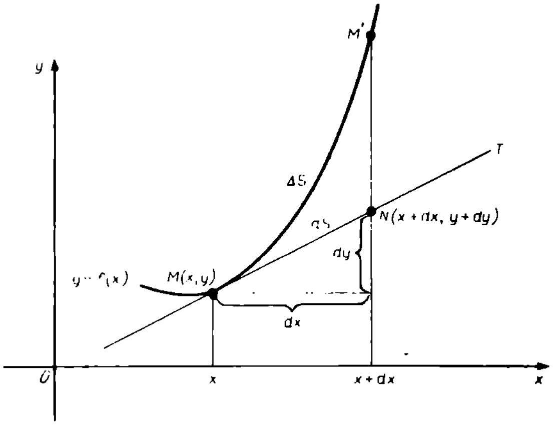


Fig. 10.29

To interpret the above formulas geometrically we look at Fig. 10.29. Clearly, the value of dS is equal to the length of the line segment MN of the tangent MT between the points $M(x, y)$ and $N(x + dx, y + dy)$.

For sufficiently small $dx = \Delta x$ the length ΔS of the arc $\overset{\smile}{MM'}$ of the curve $y = f(x)$ corresponding to the increment $\Delta x = dx$ can be thought of as being nearly equal to the length of the line segment MN of the tangent MT to the given curve at the point M, i.e., $\Delta S \simeq dS$.

When a curve is specified by the parametric equations

$$x = \varphi(t),\ y = \psi(t),\ t_0 \leqslant t \leqslant T,$$

where $\varphi(t)$ and $\psi(t)$ have continuous derivatives on $[t_0, T]$, then

$$dS = \sqrt{[\varphi'(t)]^2 + [\psi'(t)]^2}\, dt$$

or

$$dS = \sqrt{(x_t')^2 + (y_t')^2}\cdot dt.$$

If we put the parameter t equal to the length S of a variable arc, i.e., $x = \varphi(S)$ and $y = \psi(s)$, then the above formulas become

$$\left(\frac{dx}{dS}\right)^2 + \left(\frac{dy}{dS}\right)^2 = 1.$$

For a curve given by the polar equation $\varrho = f(\varphi)$, $\alpha \leqslant \varphi \leqslant \beta$, where $f(\varphi)$ is a function having a continuous derivative on the closed interval $[\alpha, \beta]$ we have

$$dS = \sqrt{\varrho^2 + (\varrho')^2}\, d\varphi.$$

10.7 Applications of the Definite Integral

Work done by a variable force. Suppose we wish to define a work done by a force **F** in moving a particle M along the x-axis from point a to point b $(a < b)$. Recall from the course of general physics that the work W done by the constant force **F** is the product of the absolute value F of this constant force by the distance $S = b - a$, i.e., $W = FS$, provided that **F** is directed along the x-axis.

Let a force **F** acting on a particle M along the x-axis be a continuous function in x ($\mathbf{F} = \mathbf{F}(x)$) on the closed interval $[a, b]$ of the x-axis. We partition the closed interval $[a, b]$ into n subintervals with lengths Δx_1, Δx_2, $\ldots$, Δx_n by choosing points in $[a, b]$ such that $x_0 = a < x_1 < x_2 < \ldots < x_n = b$. Assume that ξ_k is an arbitrary point in the subinterval $[x_{k-1}, x_k]$ and the force **F** is constant on $[x_{k-1}, x_k]$ so that $\mathbf{F} = \mathbf{F}(\xi_k)$. Then given a sufficiently small Δx_k the work ΔW_k done by $\mathbf{F}(\xi_k)$ in moving the particle from x_{k-1} to x_k is $\Delta W_k \simeq F(\xi_k)\Delta x_k$ and the sum

$$W_n = \sum_{k=1}^{n} \Delta W_k = \sum_{k=1}^{n} F(\xi_k)\Delta x_k$$

approximates the work W done by the force **F** on the closed interval $[a, b]$. Since W_n is the integral sum for the function $F(x)$ on $[a, b]$ the limit of W_n as $\max\limits_{1 \leqslant k \leqslant n} \Delta x_k \to 0$ defines the work W done by the force **F** on $[a, b]$ (this limit exists, for $F(x)$ is a continuous function on $[a, b]$). Thus the desired work W should be

$$W = \lim_{\substack{\max \Delta x_k \to 0 \\ 1 \leqslant k \leqslant n}} \sum_{k=1}^{n} F(\xi_k) \Delta x_k = \int_a^b F(x)\, dx.$$

Example. Let two point charges q_1 and q_2 be located at the points M_0 and M_1, respectively, and let r_1 be the distance between M_0 and M_1. Define the work W done in moving the charge q_2 from M_1 to M_2 assuming that the distance between M_2 and M_1 is equal to r_2 and the point M_0 is the origin of the frame of reference.

◀ Suppose that both charges are positive, i.e., $q_1 > 0$ and $q_2 > 0$; hence, they repel each other. By *Coulomb's law* the absolute value F of the force of the electrostatic interaction of two point charges in vacuum is

$$F = k \frac{q_1 q_2}{r^2},$$

where r is the distance between the charges and k is a constant.

Then applying the above formula, we obtain

$$W = \int_{r_1}^{r_2} k \frac{q_1 q_2}{r^2}\, dr = k q_1 q_2 \int_{r_1}^{r_2} \frac{dr}{r^2} = k q_1 q_2 \left(-\frac{1}{r} \right) \Bigg|_{r_1}^{r_2}$$

$$= k q_1 q_2 \left(\frac{1}{r_1} - \frac{1}{r_2} \right). \quad ▶$$

The mass and centre of gravity of a rod of varying density. Consider a rod of varying density. Think of the rod as a closed interval $[a, b]$ of the x-axis and suppose that the linear density of the rod is defined as a function $\varrho = \varrho(x)$.

We partition the closed interval $[a, b]$ into n subintervals $[x_{k-1}, x_k]$, $k = 1, 2, \ldots, n$, by choosing the points

$$x_0 < a < x_1 < x_2 < \ldots < x_{n-1} < x_n = b.$$

We let ξ_k denote an arbitrary point in the kth subinterval $[x_{k-1}, x_k]$, $k = 1, 2, \ldots, n$, and consider the sum

$$\sum_{k=1}^{n} \varrho(\xi_k) \Delta x_k.$$

Clearly, each summand gives the approximate value of the mass of the part of the rod corresponding to the subinterval $[x_{k-1}, x_k]$. Then this sum determines the approximate value of the total mass of the rod so that we can define the total mass m as the limit of $\sum_{k=1}^{n} \varrho(\xi_k)\,\Delta x_k$ as $\max\limits_{1\leqslant k\leqslant n} \Delta x_k \to 0$ which is equal to the integral $\int_{a}^{b} \varrho(x)\,dx$. Thus, the total mass m of the rod of varying density $\varrho(x)$ is given by the formula

$$m = \int_a^b \varrho(x)\,dx.$$

To define the centre of gravity of the nonhomogeneous rod we shall use the formula for the centre of mass of n particles $M_1, M_2, \ldots, M_n$ with masses $m_1, m_2, \ldots, m_n$ located at the points $x_1, x_2, \ldots, x_n$ on the x-axis. The centre of mass x_c of this system of particles is given by the formula

$$x_c = \frac{m_1x_1 + m_2x_2 + \ldots + m_nx_n}{m_1 + m_2 + \ldots + m_n} = \frac{\sum\limits_{k=1}^{n} m_kx_k}{\sum\limits_{k=1}^{n} m_k}.$$

Let us partition the closed interval $[a, b]$ into n subintervals $[x_{k-1}, x_k]$ by choosing the points $a = x_0 < x_1 < \ldots < b = x_n$ and compute the mass m_k corresponding to the kth subinterval, $k = 1, 2, \ldots, n$. Clearly, $m_k = \int_{x_{k-1}}^{x_k} \varrho(x)\,dx$. Then applying the mean value theorem (Theorem 10.5), we get

$$m_k = \varrho(\xi_k)(x_k - x_{k-1}) = \varrho(\xi_k)\,\Delta x_k,$$

where $x_{k-1} \leqslant \xi_k \leqslant x_k$.

Assume that a particle with mass m_k is located at some point ξ_k in the subinterval $[x_{k-1}, x_k]$. Then we can replace the rod by a system of n particles with masses m_k located at the points $\xi_1, \xi_2, \ldots, \xi_n$ in the closed interval $[a, b]$. Since

$$\sum_{k=1}^{n} m_k = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \varrho(x)\,dx = \int_a^b \varrho(x)\,dx = m,$$

the centre of gravity of the nonhomogeneous rod can be approximated as

$$x_c \simeq \frac{\sum\limits_{k=1}^{n} \xi_k\varrho(\xi_k)\,\Delta x_k}{m}.$$

The nominator is the integral sum for the function $x\,\varrho(x)$ on the closed interval $[a, b]$; hence, the centre of gravity of the nonhomogeneous rod is given by the formula

$$x_c = \frac{\int\limits_a^b x\,\varrho(x)\,dx}{\int\limits_a^b \varrho(x)\,dx}.$$

Example. Define the centre of gravity x_c of the nonhomogeneous rod with linear density $\varrho = x$ and length $l = 1$.

◄ The mass of the rod is

$$m = \int\limits_0^1 x\,dx = \frac{1}{2}.$$

The desired centre of gravity is

$$x_c = 2\int\limits_0^1 x^2\,dx = \frac{2}{3}.$$ ►

10.8 Numerical Integration

In dealing with physical problems we frequently work with definite integrals involving continuous functions that have no elementary antiderivatives. To compute these integrals we have to numerically approximate them. Given below are two convenient approximations, namely, *trapezoidal approximation* and *parabolic approximation*.

Trapezoidal approximation. We wish to compute the integral $\int\limits_a^b f(x)\,dx$ where $f(x)$ is a continuous function on the closed interval $[a, b]$. For simplicity we set $f(x) \geqslant 0$. Consider the partition of $[a, b]$ into n subintervals of equal length (the regular partition) by the points

$$x_0 = a < x_1 < x_2 < \ldots < x_{n-1} < x_n = b.$$

Let us draw n vertical lines $x = x_k$, $k = 0, 1, \ldots, n$ and construct n trapezoids as shown in Fig. 10.30. The total area of these n trapezoids is approximately equal to the area of the curvilinear figure $aABb$ so that

$$\int\limits_a^b f(x)\,dx \simeq \frac{f(x_0) + f(x_1)}{2}(x_1 - x_0) + \frac{f(x_1) + f(x_2)}{2}(x_2 - x_1)$$

$$+ \ldots + \frac{f(x_{n-1}) + f(x_n)}{2}(x_n - x_{n-1})$$

$$= \frac{b-a}{2n}\left[f(a) + f(b) + 2\sum_{k=1}^{n-1} f(x_k)\right],$$

where $f(x_{k-1})$ and $f(x_k)$ are the bases of the trapezoids and $x_k - x_{k-1} = \frac{b-a}{n}$ are their heights.

Thus we arrive at the formula

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2n}\left[f(a) + f(b) + 2\sum_{k=1}^{n-1} f(x_k)\right]$$

called the *trapezoidal approximation*. Notice that the larger is n the higher precision of approximation is obtained.

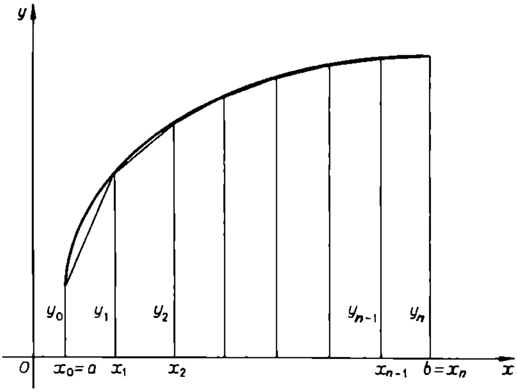


Fig. 10.30

It is worth noting that given a function $f(x)$ having the continuous second derivative $f''(x)$ on $[a, b]$ the absolute error of approximation does not exceed the number

$$M\frac{(b-a)^3}{12n^2},$$

where $M = \max\limits_{a \leqslant x \leqslant b} |f''(x)|$.

Example. Compute the trapezoidal approximation to $\int_0^1 \frac{dx}{x+1}$ with $n = 10$.

◄ Consider the regular partition of [0, 1] by the points $x_0 = 0$, $x_1 = 0.1$, $x_2 = 0.2, \ldots, x_9 = 0.9$, $x_{10} = 1$. Let us compute the values of the function $f(x) = \frac{1}{x+1}$ at these points. We have

$$\begin{array}{lll} f(0) = 1.0000, & f(0.1) = 0.9091, & f(0.2) = 0.8333, \\ f(0.3) = 0.7692, & f(0.4) = 0.7143, & f(0.5) = 0.6667, \\ f(0.6) = 0.6250, & f(0.7) = 0.5882, & f(0.8) = 0.5556, \\ f(0.9) = 0.5263, & f(1) = 0.5000. & \end{array}$$

Then the trapezoidal approximation is

$$\int_0^1 \frac{dx}{x+1} \approx \frac{1}{10}\left(\frac{1.0000 + 0.5000}{2} + 0.9091 + 0.8333 + 0.7692 \right.$$
$$\left. + 0.7143 + 0.6667 + 0.6250 + 0.5882 + 0.5556 + 0.5263\right)$$
$$= 0.69377 \approx 0.6938.$$

Let us estimate the error of the computations. Since

$$f(x) = \frac{1}{x+1} \text{ then } f'(x) = -\frac{1}{(x+1)^2} \text{ and } f''(x) = \frac{2}{(x+1)^3}.$$

On the closed interval $[0, 1]$ we have $|f''(x)| \leqslant 2$ and, consequently, $M = \max\limits_{0 \leqslant x \leqslant 1} |f''(x)| = 2$; hence the error does not exceed the number

$$M \frac{(b-a)^2}{12n^2} = \frac{2}{12 \times 10^2} = \frac{1}{600} < 0.0017.$$

The Newton-Leibniz theorem gives the exact value of this integral as

$$\int_0^1 \frac{dx}{x+1} = \ln(x+1)\Big|_0^1 = \ln 2 \simeq 0.69315.$$

The absolute error in the trapezoidal approximation is smaller than 0.0007; hence both these results are in full agreement with estimates given by theory. ►

Parabolic approximation. We start with computing the area Q of a plane figure bounded by the arc $\smile{AB}$ of the parabola $y = Ax^2 + Bx + C$ passing through the points $M_0(0, y_0)$, $M_1(h/2, y_1)$ and $M_2(h, y_2)$ (Fig. 10.31). This area is given by

$$Q = \int_0^h (Ax^2 + Bx + C)\,dx = A\frac{x^3}{3}\Big|_0^h + B\frac{x^2}{2}\Big|_0^h + Cx\Big|_0^h$$
$$= A\frac{h^3}{3} + B\frac{h^2}{2} + Ch = \frac{h}{6}(2Ah^2 + 3Bh + 6C). \qquad (*)$$

Now we express the area Q in terms of ordinates of M_1, M_2 and M_3. Substituting the coordinates of these points into the equation of the given parabola, we obtain

$$y_0 = C,\ y_1 = A\frac{h^2}{4} + B\frac{h}{2} + C,\ y_2 = Ah^2 + Bh + C.$$

Whence

$$2Ah^2 + 3Bh + 6C = y_0 + 4y_1 + y_2$$

so that

$$Q = \frac{h}{6}(y_0 + 4y_1 + y_2).$$

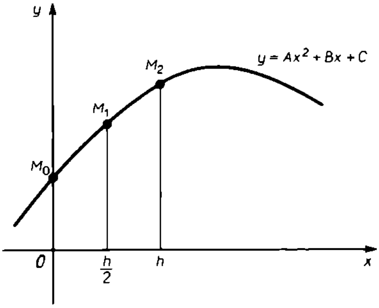


Fig. 10.31

Let us consider the definite integral $\int_a^b f(x)\,dx$ where $f(x)$ is a continuous nonnegative function on $[a, b]$. We divide $[a, b]$ into $2n$ regular subintervals (notice that $2n$ is even) by the points

$$a = x_0 < x_1 < x_2 < \ldots < x_{2n-2} < x_{2n-1} < x_{2n} = b$$

and write down the original integral in the form

$$\int_a^b f(x)\,dx = \int_{x_0}^{x_2} f(x)\,dx + \int_{x_2}^{x_4} f(x)\,dx + \ldots + \int_{x_{2n-2}}^{x_{2n}} f(x)\,dx. \qquad (**)$$

We let $A, M_1, M_2, \ldots, M_{2n-2}, M_{2n-1}, B$ denote the points of intersection of the vertical lines $x = x_k$, $k = 0, 1, 2, \ldots, 2n$, with the graph of the function $y = f(x)$; let $y_0, y_1, y_2, \ldots, y_{2n-2}, y_{2n-1}, y_{2n}$ be the ordinates of the respective intersection points. If we draw a parabola with vertical axis of symmetry through every triple of points $M_{2k-2}, M_{2k-1}, M_{2k}$ ($k = 1, 2, \ldots, n$) we get n curvilinear plane figures bounded above by the parabolas (Fig. 10.32). Since the area of the curvilinear plane figure corresponding to the subinterval $[x_{2k-2}, x_{2k}]$ with length $h = \frac{b-a}{n}$ is approximately equal to the area of the respective "parabolic" trapezoid, formula (*) gives

$$\int_{x_{2k-2}}^{x_{2k}} f(x)\,dx \simeq \frac{b-a}{6n}\,(y_{2k-2} + 4y_{2k-1} + y_{2k}),$$

where $y_k = f(x_k)$, $k = 1, 2, \ldots, n$.

Then formula (**) becomes

$$\int_a^b f(x)\,dx \simeq \frac{b-a}{6n}\,[y_0 + y_{2n} + 2(y_2 + y_4 + \ldots + y_{2n-2}) + 4(y_1 + y_3 + \ldots + y_{2n-1})].$$

This is the *parabolic approximation* which is sometimes called *Simpson's approximation.*

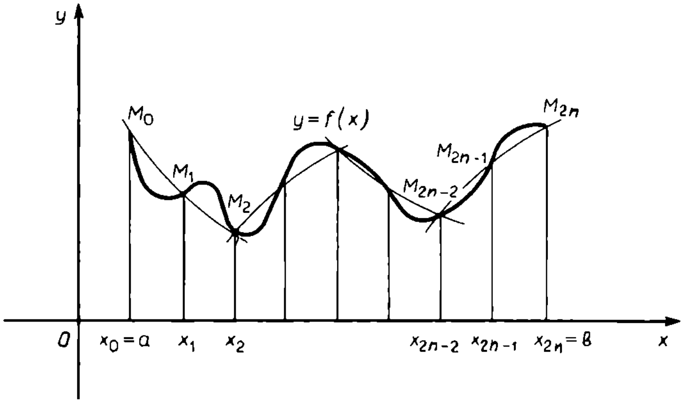


Fig. 10.32

Notice that when $f(x)$ has a continuous fourth derivative $f^{(4)}(x)$ on $[a, b]$ the error in the Simpson's approximation is at most

$$M\frac{(b-a)^5}{2880n^4},$$

where $M = \max\limits_{a \leqslant x \leqslant b} |f^{(4)}(x)|$.

The error in Simpson's approximation decreases faster than that in the trapezoidal approximation as n increases; so the former is more precise than the latter.

Example. Compute Simpson's approximation to the integral $\int_0^1 \frac{dx}{x+1}$ with $2n = 4$.

◄ Consider the regular partition of the closed interval [0, 1] by the points $x_0 = 0$, $x_1 = 0.25$, $x_2 = 0.50$, $x_3 = 0.75$, $x_4 = 1$. The values of the function $y = \dfrac{1}{x+1}$ at these points are $y_0 = 1.0000$, $y_1 = 0.8000$, $y_2 = 0.6662$, $y_3 = 0.5714$, $y_4 = 0.5000$. Then Simpson's approximation is

$$\int_0^1 \frac{dx}{x+1} \approx \frac{b-a}{6n}\,[y_0 + y_4 + 2y_2 + 4\,(y_1 + y_3)] =$$

$$\frac{1}{12}\,[1.0000 + 0.5000 + 2 \times 0.6662 + 2(0.8000 + 0.5714)] \approx 0.69325.$$

Let us estimate the error of the result thus obtained. The integrand $f(x) = \dfrac{1}{x+1}$ has the fourth derivative $f^{(4)}(x) = \dfrac{24}{(x+1)^5}$ so that

$$M = \max_{0 \leqslant x \leqslant 1} |f^{(4)}(x)| = \max_{0 \leqslant x \leqslant 1} \left| \frac{24}{(1+x)^5} \right| = 24.$$

Hence the error is at most $\dfrac{24}{2880 \times 2^4} < 0.0005$. Comparing Simpson's approximation with the exact value of the integral we conclude that the absolute error in Simpson's approximation is less than 0.0001 as it has been predicted above.

The computations of the trapezoidal approximation and Simpson's approximation to the integral $\displaystyle\int_0^1 \frac{dx}{x+1}$ prove that the latter is more precise than the former. ►

Exercises

Compute the following integrals by applying the Newton-Leibniz theorem

1. $\displaystyle\int_0^1 x^2\sqrt[5]{x}\,dx.$ **2.** $\displaystyle\int_0^4 \left(\sqrt{x} - \frac{1}{\sqrt{x}}\right) dx.$ **3.** $\displaystyle\int_e^{2e} \frac{dx}{x}.$ **4.** $\displaystyle\int_{\pi/2}^{\pi} \cos x\,dx.$

5. $\displaystyle\int_{-1/\sqrt{2}}^{1/\sqrt{2}} \frac{dx}{\sqrt{1-x^2}}.$ **6.** $\displaystyle\int_0^1 \frac{dx}{1+x^2}.$ **7.** $\displaystyle\int_{-2}^{3} |x|\,dx.$ **8.** $\displaystyle\int_0^3 \frac{x\,dx}{\sqrt{x+1}}.$

9. $\displaystyle\int_0^{-3} \frac{dx}{\sqrt{25+3x}}.$ **10.** $\displaystyle\int_e^{e^4} \frac{dx}{x\sqrt{\ln x}}.$ **11.** $\displaystyle\int_0^1 \frac{x\,dx}{\sqrt{1-x^2}}.$

12. $\int\limits_0^4 \frac{\ln(x+\sqrt{9+x^2})}{\sqrt{9+x^2}}\,dx.$ **13.** $\int\limits_0^2 \frac{x\,dx}{\sqrt{1+x^4}}.$ **14.** $\int\limits_0^1 \frac{x\,dx}{\sqrt{4-x^4}}.$

15. $\int\limits_0^{\pi/2} 3^{\cos^2 x}\sin 2x\,dx.$ **16.** $\int\limits_0^{\pi/2} \frac{\cos x-\sin x}{\cos x+\sin x}\,dx.$

17. $\int\limits_0^{\pi/12} \ln\sin 2x\cdot\cot 2x\,dx.$ **18.** $\int\limits_1^{e} \frac{\cos(\ln x)}{x}\,dx.$ **19.** $\int\limits_0^1 \sinh^2 x\,dx.$

20. $\int\limits_0^{\ln 2} \tanh x\,dx.$ **21.** $\int\limits_0^{\ln 3} \tanh^2 x\,dx.$

Integrate by parts each of the following integrals

22. $\int\limits_0^{2\pi} x\sin x\,dx.$ **23.** $\int\limits_0^1 \ln(1+x)\,dx.$ **24.** $\int\limits_0^1 x\sinh x\,dx$

25. $\int\limits_0^{1/\sqrt{2}} \sin^{-1} x\,dx.$ **26.** $\int\limits_0^1 x^3 e^{x^2/2}\,dx.$

Compute the areas of the plane figures bounded by the given curves and the indicated axes

27. $y = x^2 + 2x - 3$ (parabola), $y = x + 3$ (line).
28. $y = 2x - x^2$ (parabola), $y = -x$ (line).
29. $y = x^2 - 1$ (parabola), $x = 2$ (line); the x- and y-axes.
30. $y = x^2 - 3x - 4$, $y = 4 + 3x - x^2$ (parabolas).
31. $y = x - 1$, $y = 1$, $y = \ln x$.
32. $y = e^{-x}$, $y = e^x$, $x = 1$.
33. $y = x^3$, $y = 8$; the y-axis.
34. $x = a\cos^3 t$ and $y = a\sin^3 t$, $a > 0$ (astroid).
35. $x = a(t - \sin t)$ and $y = a(1 - \cos t)$, $a > 0$ (an arch of a cycloid); the axis of abscissas.
36. $x = a(2\cos t - \cos 2t)$ and $y = a(2\sin t - \sin 2t)$, $a > 0$ (cardioid).
37. $\varrho = a\sin\varphi$, $a > 0$. **38.** $\varrho = a\sin 2\varphi$, $a > 0$. **39.** $\varrho = 2 + \sin\varphi$.

Compute the volume of the solid generated by revolving the given curves or the given plane figures around the indicated axis

40. $y = \sin x$, $0 \leqslant x \leqslant \pi$; around the x-axis.
41. $y = \sin^2 x$, $0 \leqslant x \leqslant \pi$; around the x-axis.
42. $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ (ellipse); around the x-axis.

43. The plane figure bounded by the parabola $y = ax - x^2$, $a > 0$, and the x-axis; around the x-axis.

44. The plane figure bounded by the parabola $y = x^2$, the line $y = 1$ and the y-axis; around the y-axis.

45. Compute the volume of the solid bounded by the paraboloid $\frac{y^2}{2p} + \frac{z^2}{2q} = x$ and the plane $x = a$.

46. Compute the volume of the solid bounded by the hyperboloid of one sheet $\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1$ and the planes $z = 0$ and $z = h$.

Compute the length of the arc of the given curve between the indicated points

47. $y = \frac{x^2}{2}$; from $(0, 0)$ to $(1, 1/2)$.

48. $y = \sqrt{x^3}$ (semicubical parabola); from the origin of coordinates to $A(1, 1)$.

49. $y = \ln x$; from $x = \sqrt{3}$ to $x = \sqrt{8}$.

50. $y = \ln \sin x$; from $x = \pi/3$ to $x = \pi/2$.

51. $x = a \cos^3 t$ and $y = a \sin^3 t$, $a > 0$ (astroid; define the total length).

52. $x = \frac{t^3}{3} - t$ and $y = t^2 + 2$; from $t = 0$ to $t = 3$.

53. $x = e^t \cos t$ and $y = e^t \sin t$; from $t = 0$ to $t = \ln \pi$.

54. $\varrho = a e^{\varphi}$, $a > 0$, (logarithmic spiral); inside the circle $\varrho \leqslant a$.

55. $\varrho = a \sin \varphi$, $a > 0$; (total length).

56. $\varrho = a\varphi$, $a > 0$, (spiral of Archimedes); the length of the first winding.

Answers

1. $\frac{5}{17}$. **2.** $\frac{4}{3}$. **3.** $\ln 2$. **4.** -1. **5.** $\frac{\pi}{2}$. **6.** $\frac{\pi}{4}$. **7.** 6.5. **8.** $\frac{8}{3}$. **9.** $-\frac{2}{3}$. **10.** 2. **11.** 1. **12.** $\frac{3}{2} \ln^2 3$. **13.** $\frac{1}{2} \ln (4 + \sqrt{17})$. **14.** $\frac{\pi}{4}$. **15.** $\frac{2}{\ln 3}$. **16.** 0. **17.** $-\frac{\ln 2}{4}$. **18.** $\sin 1$. **19.** $\sinh^2 1$. **20.** $-\ln \frac{5}{4}$. **21.** $\ln 3 - \frac{4}{5}$. **22.** -2π. **23.** $\ln 4 - 1$. **24.** e^{-1}. **25.** $\frac{\pi + 4 - \sqrt{2}}{2}$. **26.** $2 - \sqrt{e}$. **27.** $\frac{125}{6}$. **28.** 4.5. **29.** 2. **30.** $\frac{125}{3}$. **31.** $e - 2.5$. **32.** $2 \cosh 1 - 2$. **33.** 12. **34.** $\frac{3}{8} \pi a^2$. **35.** $3\pi a^2$. **36.** $6\pi a^2$. **37.** $\frac{\pi a^2}{2}$. **38.** $\frac{\pi a^2}{4}$. **39.** $\frac{9}{2} \pi$. **40.** $\frac{\pi^2}{2}$. **41.** $\frac{3}{8} \pi^2$. **42.** $\frac{4}{3} \pi a b^2$. **43.** $\frac{\pi a^5}{30}$. **44.** $\frac{\pi}{2}$. **45.** $\pi a^2 \sqrt{pq}$. **46.** $\pi abh \left(1 + \frac{h^2}{3C^2}\right)$. **47.** $\frac{1}{2} [\sqrt{2} + \ln (1 + \sqrt{2})]$. **48.** $\frac{13\sqrt{13} - 8}{27}$. **49.** $1 + \frac{1}{2} \ln \frac{3}{2}$. **50.** $\frac{1}{2} \ln 3$. **51.** $6a$. **52.** 12. **53.** $(\pi - 1)\sqrt{2}$. **54.** $a\sqrt{2}$. **55.** πa.

56. $\pi a \sqrt{1 + 4\pi^2} + \frac{a}{2} \ln \left((2\pi + \sqrt{1 + 4\pi^2}\right)$.

Chapter 11

Improper Integrals

11.1 Integrals with Infinite Limits of Integration

The notion of the definite integral involves a function which is defined on a closed (bounded) interval $[a, b]$ so that the interval of integration is always bounded. Certain applications of integral calculus lead to integrals of functions defined over unbounded domains; these are infinite semi-intervals of forms $[a, +\infty)$, and $(-\infty, b]$ or the infinite interval of the form $(-\infty, +\infty)$. For example, we encounter such integrals when computing the potential of gravitational or electrostatic forces.

To make the notion of the definite integral applicable to unbounded intervals of integration we must rely upon some new concepts and definitions which clarify the meanings of the symbols

$$\int\limits_a^{+\infty} f(x)\,dx, \quad \int\limits_{-\infty}^{b} f(x)\,dx, \quad \int\limits_{-\infty}^{+\infty} f(x)\,dx.$$

Let a function $f(x)$ be defined for all $x \geqslant a$ and integrable, say continuous, on every finite closed interval $a \leqslant x \leqslant b$ where a is a given number and b ($b \geqslant a$) is any arbitrary number. To deduce what we mean by writing

$$\int\limits_a^{+\infty} f(x)\,dx \tag{*}$$

which is called the *improper integral*, we consider the function $J(b) = \int\limits_a^b f(x)\,dx$ of the variable b ($b \geqslant a$).

Definition. If the function $J(b)$ has a finite limit L as $b \to +\infty$ the improper integral (*) is said to *converge*.

In this case we write

$$\int\limits_a^{+\infty} f(x)\,dx = \lim_{b \to +\infty} \int\limits_a^b f(x)\,dx = L.$$

Suppose that the function $J(b)$ has no (finite) limit as $b \to +\infty$. Then the improper integral (*) is said to *diverge*; in this case no numerical value is assigned to this integral.

Examples. (1) The improper integral $\int\limits_0^{+\infty} \frac{dx}{1+x^2}$ converges and is equal to $\frac{\pi}{2}$.

◀ Indeed, by the definition we have

$$\int\limits_0^{+\infty} \frac{dx}{1+x^2} = \lim_{b\to+\infty} \int\limits_0^{b} \frac{dx}{1+x^2} = \lim_{b\to+\infty} \tan^{-1} b = \frac{\pi}{2}. \quad ▶$$

(2) The improper integral $\int\limits_0^{+\infty} \cos x\, dx$ diverges.

◀ Indeed, the integral $\int\limits_0^{b} \cos x\, dx = \sin b$ has no limit as $b \to +\infty$; hence, it diverges. ▶

(3) Let two charges q_1 and q_2 be both positive so that they repel each other. By the Coulomb's law the absolute value F of the force of electrostatic interaction of two point charges in vacuum is given by

$$F = \frac{kq_1q_2}{r^2},$$

where r is the distance between the charges and k is a constant.

Suppose that the charge q_1 is located at the point M_0 which is chosen as the origin of some reference frame and the charge q_2 is located at the point M. We let r_1 denote the distance between the ponts M and M_0 and compute the work done in moving the charge q_2 from M to infinity.

◀ The desired work W is defined as the improper integral

$$W = \int\limits_{r_1}^{+\infty} \frac{kq_1q_2}{r^2}\, dr = kq_1q_2 \int\limits_{r_1}^{+\infty} \frac{dr}{r^2}.$$

By the definition we have

$$\int\limits_{r_1}^{+\infty} \frac{dr}{r^2} = \lim_{b\to+\infty} \int\limits_{r_1}^{b} \frac{dr}{r^2} = \lim_{b\to+\infty} \left(-\frac{1}{r}\right)\Bigg|_{r=r_1}^{r=b} = \lim_{b\to+\infty} \left(\frac{1}{r_1} - \frac{1}{b}\right) = \frac{1}{r_1}.$$

Therefore, $W = \frac{kq_1q_2}{r_1}$. Let q_2 be a unit charge. Then $W = \frac{kq_1}{r_1}$; this quantity is called the *potential of the field induced by the charge* q_1. ▶

(4) Consider the integral $\int\limits_1^{+\infty} \frac{dx}{x^\alpha}$, $\alpha = \text{const}$.

We wish to determine the values of α for which the integral converges and those for which it diverges.

◄ By the definition of the improper integral we have

$$\int_1^{+\infty} \frac{dx}{x^\alpha} = \lim_{b \to +\infty} \int_1^b \frac{dx}{x^\alpha}.$$

Let $\alpha \neq 1$. Then

$$\int_1^b \frac{dx}{x^\alpha} = \frac{x^{1-\alpha}}{1-\alpha}\Bigg|_1^b = \frac{b^{1-\alpha}}{1-\alpha} - \frac{1}{1-\alpha}$$

so that for $\alpha > 1$ we have

$$\lim_{b \to +\infty} \int_1^b \frac{dx}{x^\alpha} = \lim_{b \to +\infty} \left(\frac{b^{1-\alpha}}{1-\alpha} - \frac{1}{1-\alpha} \right) = \frac{1}{\alpha - 1}.$$

Thus the original integral converges if $\alpha > 1$. For $\alpha < 1$ the integral $\int_1^b \frac{dx}{x^\alpha}$ has no finite limit as $b \to +\infty$; in other words the integral diverges if $\alpha < 1$.

Now let $\alpha = 1$. Then

$$\int_1^b \frac{dx}{x^\alpha} = \int_1^b \frac{dx}{x} = \ln b.$$

Whence it follows that for $\alpha = 1$ $\int_1^b \frac{dx}{x^\alpha} \to +\infty$ as $b \to +\infty$.

Therefore we infer that

$$\text{the integral } \int_1^{+\infty} \frac{dx}{x^\alpha} \begin{cases} \text{converges} & \text{if } \alpha > 1, \\ \text{diverges} & \text{if } \alpha \leqslant 1. \end{cases}$$

This result can easily be interpreted geometrically. Consider the region D bounded from the left by the line $x = 1$, below by the x-axis and above by the graph of the function $y = 1/x^\alpha$ (Fig. 11.1). This region is unbounded from the right, i.e., it extends to infinity. By convention we regard the limit of the region bounded from the right by the line $x = b$ provided that $b \to +\infty$ is the area of the infinite region D (Fig. 11.2). It is easy to see that for $\alpha > 1$ the area of the region D bounded above by the graph of $y = 1/x^\alpha$ is finite; if D is bounded above by the graph of the hyperbola $y = 1/x$ or of the curve $y = 1/x^\alpha$ with $\alpha < 1$ the notion of the area of D has no meaning.

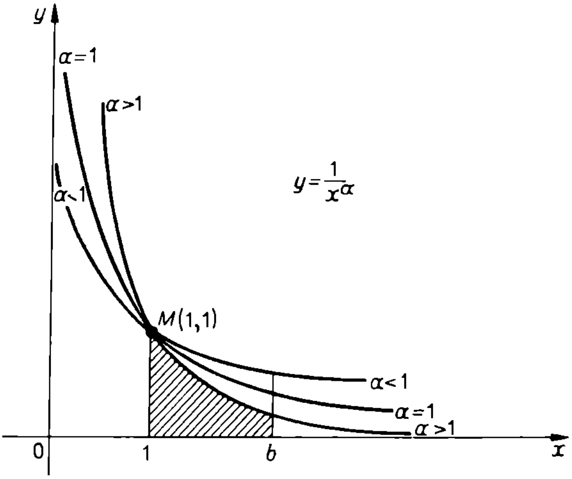


Fig. 11.1

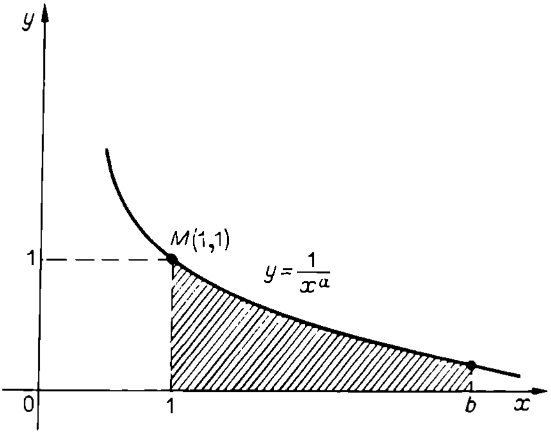


Fig. 11.2

Remark. Notice that given any $a \geqslant \delta > 0$ the integral $\int_a^{+\infty} \frac{dx}{x^\alpha}$ converges for $\alpha > 1$ and diverges for $\alpha \leqslant 1$.

The definition of the improper integral

$$\int_a^{+\infty} f(x)\,dx = \lim_{b\to+\infty} \int_a^b f(x)\,dx$$

implies directly the following

(1) If the integral $\int\limits_a^{+\infty} f(x)\,dx$ converges so does the integral $\int\limits_a^{+\infty} \lambda f(x)\,dx$ where λ is any real number and

$$\int\limits_a^{+\infty} \lambda f(x)\,dx = \lambda \int\limits_a^{+\infty} f(x)\,dx.$$

(2) If both $\int\limits_a^{+\infty} f(x)\,dx$ and $\int\limits_a^{+\infty} \varphi(x)\,dx$ converge so does $\int\limits_a^{+\infty} (f(x) + \varphi(x))\,dx$ and

$$\int\limits_a^{+\infty} (f(x) + \varphi(x))\,dx = \int\limits_a^{+\infty} f(x)\,dx + \int\limits_a^{+\infty} \varphi(x)\,dx.$$

◀ Indeed, given any $b > a$ there holds

$$\int\limits_a^{b} (f(x) + \varphi(x))\,dx = \int\limits_a^{b} f(x)\,dx + \int\limits_a^{b} \varphi(x)\,dx.$$

Both integrals on the right have limits as $b \to +\infty$; hence, the integral on the left also has a limit as $b \to +\infty$, i.e., $\int\limits_a^{+\infty} (f(x) + \varphi(x))\,dx$ converges. Evaluating the limit of the above identity as $b \to +\infty$, we obtain the desired result. ▶

Problem. Let the integral $\int\limits_a^{+\infty} (f(x) + \varphi(x))\,dx$ converge. Determine whether or not the improper integrals $\int\limits_a^{+\infty} f(x)\,dx$ and $\int\limits_a^{+\infty} \varphi(x)\,dx$ converge.

(3) Let $u(x)$ and $v(x)$ be continuously differentiable functions on the line $x \geqslant a$. Then

$$\int\limits_a^{+\infty} u\,dv = (u(x)\,v(x))\Big|_a^{+\infty} - \int\limits_a^{+\infty} v\,du \quad \text{(integration by parts)}$$

provided that at least two of the three terms involved make sense (notice that in this case the third term also makes sense).

Example. Integrate by parts $J_n = \int\limits_0^{+\infty} x^n e^{-x}\,dx$, where n is a natural number or zero.

◄ We have

$$J_n = \int\limits_0^{+\infty} x^n e^{-x}\,dx = -(x^n e^{-x})\Big|_0^{+\infty} + n\int\limits_0^{+\infty} x^{n-1}e^{-x}\,dx = nJ_{n-1},$$

$n = 1, 2, \ldots$.

Notice that

$$J_0 = \int\limits_a^{+\infty} e^{-x}\,dx = 1$$

we get $J_n = n!$ ►

11.2 Integrals of Nonnegative Functions

In a variety of applications it suffices to determine whether a given integral converges or not without having to evaluate it directly. Convergence tests are convenient devices in many instances. Here we formulate some valuable convergence tests for improper integrals in the forms of theorems.

Theorem 11.1. *Let $f(x)$ and $\varphi(x)$ be integrable functions on $[a, b]$ for any $b > a$ and let $0 \leqslant f(x) \leqslant \varphi(x)$ for all $x \geqslant a$. Then*

(1) *If $\int\limits_a^{+\infty} \varphi(x)\,dx$ converges so does $\int\limits_a^{+\infty} f(x)\,dx$.*

(2) *If $\int\limits_a^{+\infty} f(x)\,dx$ diverges so does $\int\limits_a^{+\infty} \varphi(x)\,dx$.*

◄ First we suppose that the integral $\int\limits_a^{+\infty} \varphi(x)\,dx$ converges and prove that so does $\int\limits_a^{+\infty} f(x)\,dx$. In other words we have to show that the function $J(b) = \int\limits_a^{b} f(x)\,dx$ has a finite limit as $b \to +\infty$.

Since $f(x)$ is a nonnegative function for all $x \geqslant a$ $J(b)$ must be a nondecreasing function in b. Indeed, if $b_1 > b$ then

$$J(b_1) = \int\limits_a^{b_1} f(x)\,dx = \int\limits_a^{b} f(x)\,dx + \int\limits_b^{b_1} f(x)\,dx \geqslant \int\limits_a^{b} f(x)\,dx = J(b).$$

Also, since $f(x) \leqslant \varphi(x)$ for all $x \geqslant a$ then given any $b > a$ there holds

$$\int\limits_a^{b} f(x)\,dx \leqslant \int\limits_a^{b} \varphi(x)\,dx.$$

The value of the integral $\int_a^b \varphi(x)\,dx$ is not larger than that of the integral $\int_a^{+\infty} \varphi(x)\,dx$; the latter integral converges by the hypothesis. Hence, given any $b > a$ there holds

$$J(b) = \int_a^b f(x)\,dx \leqslant \int_a^{+\infty} \varphi(x)\,dx = L$$

so that $J(b) = \int_a^b f(x)\,dx$ is a nondecreasing function in b which is bounded above (as $b \to +\infty$). This means that $J(b)$ has a finite limit as $b \to +\infty$ and by virtue of the definition the integral $\int_a^{+\infty} f(x)\,dx$ converges.

Now we prove the second part of the theorem. Let the integral $\int_a^{+\infty} f(x)\,dx$ diverge. Assume the converse, i.e., assume that the integral $\int_a^{+\infty} \varphi(x)\,dx$ converges. Then the first part of the theorem tells that the integral $\int_a^{+\infty} f(x)\,dx$ must converge and that contradicts to the hypothesis of the theorem. Hence, our assumption is false and the integral $\int_a^{+\infty} \varphi(x)\,dx$ diverges. ►

Example. Investigate the improper integral $\int_0^{+\infty} \frac{e^{-x^2}}{1 + x^2 + \sin^4 x}\,dx$.

◄ Notice that for all $x \geqslant 0$ the function $f(x) = \frac{e^{-x^2}}{1 + x^2 + \sin^4 x}$ is such that

$$0 < \frac{e^{-x^2}}{1 + x^2 + \sin^4 x} \leqslant \frac{1}{1 + x^2} = \varphi(x).$$

The integral $\int_0^{+\infty} \frac{dx}{1 + x^2}$ converges (See Example (1) on p. 507). Then by virtue of part (1) of Theorem 11.1 the integral in question also converges. It is worth mentioning here that we fail to investigate this integral by applying the definition of convergent integrals. ►

Theorem 11.2. *Let $f(x)$ and $\varphi(x)$ be continuous nonnegative functions for all $x \geqslant a$ and let $\varphi(x)$ be distinct from zero for all sufficiently large x. Then both integrals $\int\limits_a^{+\infty} f(x)\,dx$ and $\int\limits_a^{+\infty} \varphi(x)\,dx$ either converge or diverge provided that there exists*

$$\lim_{x\to+\infty} \frac{f(x)}{\varphi(x)} = k \neq 0.$$

◄ Indeed, let $\lim\limits_{x\to+\infty} \dfrac{f(x)}{\varphi(x)} = k > 0$. By the definition of the limit this means that given any number $\varepsilon > 0$, say $\varepsilon = \dfrac{k}{2} > 0$, there exists a number N such that for all $x \geqslant N$

$$\left| \frac{f(x)}{\varphi(x)} - k \right| < \varepsilon = \frac{k}{2}$$

or, equivalently,

$$\frac{k}{2} < \frac{f(x)}{\varphi(x)} < \frac{3}{2}\,k.$$

Recall that $\varphi(x) > 0$; then the above inequality becomes

$$\frac{k}{2}\,\varphi(x) < f(x) < \frac{3}{2}\,k\varphi(x) \qquad \forall x \geqslant N.$$

Theorem 11.1 implies that

(a) If $\int\limits_N^{+\infty} \varphi(x)\,dx$ converges so does $\int\limits_N^{+\infty} f(x)\,dx$ since $f(x) < \dfrac{3}{2}\,k\varphi(x)$,

(b) If $\int\limits_N^{+\infty} \varphi(x)\,dx$ diverges so does $\int\limits_N^{+\infty} f(x)\,dx$ since $\dfrac{k}{2}\,\varphi(x) < f(x)$.

Proceeding in a similar way we infer that if $\int\limits_N^{+\infty} f(x)\,dx$ converges (diverges) then $\int\limits_N^{+\infty} \varphi(x)\,dx$ also converges (diverges).

These results remain valid for $\int\limits_a^{+\infty} f(x)\,dx$ and $\int\limits_a^{+\infty} \varphi(x)\,dx$ due to the fact that the integral $\int\limits_a^{+\infty} g(x)\,dx$ converges or diverges simultaneously with the integral $\int\limits_p^{+\infty} g(x)\,dx$, where $p > a$ is any whatever large fixed number since the difference of these integrals is also an improper integral. ►

Example. Determine whether or not the improper integral

$$\int_1^{+\infty} \frac{2x^2+1}{x^3+3x+4}\,dx$$

converge.

◀ The integrand $f(x) = \dfrac{2x^2+1}{x^3+3x+4}$ is positive for all $x \geqslant 1$. Write down $f(x)$ as

$$f(x) = \frac{2+1/x^2}{x+3/x+4/x^2}.$$

Whence we conclude that for large x the function $f(x)$ is similar to the function $2/x$.

Put $\varphi(x) = \dfrac{1}{x}$ and consider the integral $\displaystyle\int_1^{+\infty} \frac{dx}{x} = \int_1^{+\infty} \varphi(x)\,dx$. This integral diverges and by virtue of Theorem 11.2 so does the original integral since

$$\lim_{x\to+\infty} \frac{f(x)}{\varphi(x)} = \lim_{x\to+\infty} \frac{(2x^2+1)x}{x^3+3x+4} = 2 \neq 0. \ ▶$$

Theorem 11.1 along with the results concerning the integral $\displaystyle\int_1^{+\infty} \frac{dx}{x^\alpha}$ lead to the following convergence test for the integral $\displaystyle\int_a^{+\infty} f(x)\,dx$ involving a nonnegative function $f(x)$.

Theorem 11.3. *Let there exist a number $\alpha > 1$ such that for all sufficiently large x*

$$0 \leqslant f(x) \leqslant \frac{M}{x^\alpha},$$

where M is independent of x and $M > 0$. Then the integral $\displaystyle\int_a^{+\infty} f(x)\,dx$ converges.

If $f(x) \geqslant \dfrac{M}{x}$ where M is independent of x and $M > 0$, whenever x is sufficiently large, then the integral $\displaystyle\int_a^{+\infty} f(x)\,dx$ diverges.

◀ Let $0 \leqslant f(x) \leqslant \dfrac{M}{x^\alpha} (\alpha > 1)$ for all $x \geqslant A > \max\{a, 0\}$. Since the integral $\displaystyle\int_A^{+\infty} \frac{M}{x^\alpha}\,dx$ where $\alpha > 1$ converges then putting $\varphi(x) = \dfrac{M}{x^\alpha}$ in Theorem 11.1, we conclude that so does $\int_A^{+\infty} f(x)\,dx$. Whence it follows that $\int_a^{+\infty} f(x)\,dx$ converges, for

$$\int_a^b f(x)\,dx = \int_a^A f(x)\,dx + \int_A^b f(x)\,dx$$

and the integrals $\int_a^b f(x)\,dx$ and $\int_A^b f(x)\,dx$ must simultaneously have finite limits as $b \to +\infty$.

Now let $f(x) \geqslant \frac{M}{x}$ $(M > 0)$ for all $x \geqslant A > \max\{a, 0\}$. Since the integral $\int_A^{+\infty} \frac{M}{x} dx$ diverges then, by virtue of Theorem 11.1, so does the integral $\int_A^{+\infty} f(x)\,dx$; this implies that the integral $\int_a^{+\infty} f(x)\,dx$ also diverges. ►

Example. Investigate the integral $\int_3^{+\infty} \frac{x-2}{x^3 + x^2 + 2x + 5}\,dx$.

◄ For $x \geqslant 3$ we have

$$0 < \frac{x-2}{x^3 + x^2 + 2x + 5} < \frac{x}{x^3} = \frac{1}{x^2}.$$

The integral $\int_3^{+\infty} \frac{dx}{x^2}$ converges since $\alpha = 2 > 1$; hence so does the original integral. ►

11.3 Absolutely Convergent Improper Integrals

Let a function $f(x)$ be defined for all $x \geqslant a$ and integrable on every $[a, b]$ where $b > a$.

Definition. The integral $\int_a^{+\infty} f(x)\,dx$ is said to *converge absolutely* if the integral $\int_a^{+\infty} |f(x)|\,dx$ converges.

If the integral $\int_a^{+\infty} f(x)\,dx$ converges and the integral $\int_a^{+\infty} |f(x)|\,dx$ diverges we say that the integral $\int_a^{+\infty} f(x)\,dx$ is *conditionally convergent.*

Theorem 11.4. *If the integral* $\int_a^{+\infty} |f(x)|\, dx$ *converges so does the integral* $\int_a^{+\infty} f(x)\, dx.$

◀ Let $\int_a^{+\infty} |f(x)|\, dx$ converge, i.e.,

$$\lim_{b\to+\infty} \int_a^b |f(x)|\, dx = L < +\infty.$$

Since given any x in the domain of $f(x)$ there holds

$$-|f(x)| \leqslant f(x) \leqslant |f(x)|$$

then

$$0 \leqslant |f(x)| + f(x) \leqslant 2|f(x)|. \tag{*}$$

The integral $\int_a^{+\infty} |f(x)|\, dx$ converges by the hypothesis; hence, so does the integral $\int_a^{+\infty} 2|f(x)|\, dx = 2\int_a^{+\infty} |f(x)|\, dx$. Then by virtue of (*) Theorem 11.1 yields that the integral $\int_a^{+\infty} (f(x) + |f(x)|)\, dx$ converges. This means that the integral $\int_a^b (f(x) + |f(x)|)\, dx$ has a finite limit as $b \to +\infty$.

Evidently, we have for all $x \geqslant a$

$$f(x) = (f(x) + |f(x)|) - |f(x)|.$$

Whence for any $b > a$ there holds

$$\int_a^b f(x)\, dx = \int_a^b (f(x) + |f(x)|)\, dx - \int_a^b |f(x)|\, dx.$$

Both integrals on the right have finite limits as $b \to +\infty$. Hence, the integral $\int_a^b f(x)\, dx$ also has a finite limit as $b \to +\infty$, i.e., the integral $\int_a^{+\infty} f(x)\, dx$ converges. ▶

Theorem 11.1 and the results concerning the integrals of the form $\int_1^{+\infty} \frac{dx}{x^\alpha}$ leads to the following convergence test for the integral $\int_a^{+\infty} f(x)\, dx$.

Theorem 11.5. *Let there exist a number* $\alpha > 1$ *such that for all sufficiently large* x *there holds*

$$|f(x)| \leqslant \frac{M}{x^\alpha},$$

where M is independent of x and $M > 0$. Then the integral $\int_a^{+\infty} f(x)\,dx$ converges absolutely.

◀ Suppose that the above condition holds for all $x \geqslant A > \max\{a, 0\}$. Since the integral $\int_A^{+\infty} \frac{M}{x^\alpha}\,dx$ where $\alpha > 1$ converges, then by virtue of Theorem 11.1 so does the integral $\int_A^{+\infty} |f(x)|\,dx$. This implies that the integral $\int_a^{+\infty} |f(x)|\,dx$ converges and, consequently, the integral $\int_a^{+\infty} f(x)\,dx$ converges absolutely.

If $|f(x)| \geqslant \frac{M}{x^\alpha}$ the integral $\int_a^{+\infty} f(x)\,dx$ does not converge absolutely. ▶

For example, the integral $\int_1^{+\infty} \frac{\sin x}{x^2}\,dx$ converges absolutely since $\left|\frac{\sin x}{x^2}\right| \leqslant \frac{1}{x^2}$ for all x, $x \geqslant 1$, and the integral $\int_1^{+\infty} \frac{dx}{x^2}$ converges.

Thus we infer that the integral $\int_a^{+\infty} f(x)\,dx$ converges if it converges absolutely. The converse is not true: not every convergent integral converges absolutely. For example, the integral $\int_1^{+\infty} \frac{\sin x}{x}\,dx$ being convergent does not converge absolutely.

◀ Indeed, applying integration by parts, we obtain

$$\int_1^{+\infty} \frac{\sin x}{x}\,dx = -\int_1^{+\infty} \frac{1}{x}\,d(\cos x) = -\left.\frac{\cos x}{x}\right|_1^{+\infty} - \int_1^{+\infty} \frac{\cos x}{x^2}\,dx$$

$$= \cos 1 - \int_1^{+\infty} \frac{\cos x}{x^2}\,dx.$$

The integral $\int_1^{+\infty} \frac{\cos x}{x^2}\,dx$ converges absolutely; hence, it is convergent.

Thus both expressions on the right of the above identity are finite so that (1) the method of integration by parts is justified and (2) the integral on the left is finite, i.e., the integral $\int_1^{+\infty} \frac{\sin x}{x}\,dx$ converges.

Now we show that the integral $\int\limits_{1}^{+\infty} \frac{\sin x}{x}\,dx$ is conditionally convergent. In other words we show that the integral $\int\limits_{1}^{+\infty} \frac{|\sin x|}{x}\,dx$ diverges.

Notice that for any $b > 1$ the inequality

$$|\sin x| \geqslant \sin^2 x = \frac{1 - \cos 2x}{2}$$

implies that

$$\int\limits_{1}^{b} \frac{|\sin x|}{x}\,dx \geqslant \frac{1}{2} \int\limits_{1}^{b} \frac{dx}{x} - \frac{1}{2} \int\limits_{1}^{b} \frac{\cos 2x}{x}\,dx. \qquad (*)$$

The integral $\int\limits_{1}^{+\infty} \frac{dx}{x}$ diverges, for $\lim\limits_{b \to +\infty} \int\limits_{1}^{b} \frac{dx}{x} = +\infty$. Applying integration by parts to the integral $\int\limits_{1}^{\infty} \frac{\cos 2x}{x}\,dx$, we infer that it converges. Evaluating the limit of (*) as $b \to +\infty$, we conclude that the right-hand side of (*) and, hence, the left-hand side of (*) diverge to infinity. This means that the integral $\int\limits_{1}^{+\infty} \frac{|\sin x|}{x}\,dx$ diverges, whence it follows that the original integral $\int\limits_{1}^{+\infty} \frac{\sin x}{x}\,dx$ is not absolutely convergent. ►

Now we turn our attention to *Dirichlet's convergence test* which specifies the sufficient conditions for the improper integral to converge.

Theorem 11.6 (Dirichlet's convergence test). *If a function is continuous and has a bounded antiderivative $F(x)$ for $x \geqslant a$ and if a function $g(x)$ is continuously differentiable and monotonically decreases for $x \geqslant a$ and if $\lim\limits_{x \to +\infty} g(x) = 0$, then the integral $\int\limits_{a}^{+\infty} f(x)\,g(x)\,dx$ converges.*

The proof of this theorem can easily be found elsewhere in the relevant literature.

By way of illustration we apply Dirichlet's convergence test to the integral $\int\limits_{1}^{+\infty} \frac{\sin x}{x^{\alpha}}\,dx$, $\alpha > 0$.

◀ The function $f(x) = \sin x$ has the bounded antiderivative $F(x) = -\cos x$ for all x; $g(x) = 1/x^{\alpha}$ $(\alpha > 0)$ is a continuously differentiable function for $x \geqslant 1$ and $g(x)$ monotonically decreases and tends to zero as $x \to \infty$. Hence, all the conditions of Theorem 11.6 are satisfied so that the integral in question converges. ▶

Problem. Show that the Fresnel integral $\int\limits_{0}^{+\infty} \sin x^2\, dx$ converges. (*Hint.* Use the substitution $x^2 = t$.)

11.4 Cauchy Principal Value of the Improper Integral

We define the improper integral $\int\limits_{-\infty}^{b} f(x)\, dx$ by writing

$$\int\limits_{-\infty}^{b} f(x)\, dx = \lim_{a \to -\infty} \int\limits_{a}^{b} f(x)\, dx$$

and say that $\int\limits_{-\infty}^{b} f(x)\, dx$ converges if this limit exists; otherwise the improper integral is said to diverge.

If both the lower and upper limits of integration are infinite we set

$$\int\limits_{-\infty}^{+\infty} f(x)\, dx = \lim_{\substack{b \to +\infty \\ a \to -\infty}} \int\limits_{a}^{b} f(x)\, dx$$

or

$$\int\limits_{-\infty}^{+\infty} f(x)\, dx = \lim_{\substack{N_1 \to +\infty \\ N_2 \to +\infty}} \int\limits_{-N_1}^{N_2} f(x)\, dx,$$

where N_1 and N_2 diverge to $+\infty$ independently. It may happen that the improper integral thus defined has no meaning while there exists its *Cauchy principal value* (C.p.v.) given by the formula

$$\text{C.p.v.} \int\limits_{-\infty}^{+\infty} f(x)\, dx = \lim_{N \to +\infty} \int\limits_{-N}^{N} f(x)\, dx,$$

i.e., when $N_1 = N_2 = N$. Then we say that the improper integral $\int\limits_{-\infty}^{+\infty} f(x)\, dx$ converges in the sense of the Cauchy principal value.

Example. Investigate the integral $\displaystyle\int\limits_{-\infty}^{+\infty} \frac{x\, dx}{1 + x^2}$.

◀ We have $\int\limits_{N_1}^{N_2} \frac{x\,dx}{1+x^2} = \frac{1}{2}\ln\frac{1+N_2^2}{1+N_1^2}$, whence it follows that the integral $\int\limits_{-N_1}^{N_2} \frac{x\,dx}{1+x^2}$ has no limit as $N_1 \to +\infty$ and $N_2 \to +\infty$ so that the integral $\int\limits_{-\infty}^{+\infty} \frac{x\,dx}{1+x^2}$ diverges. On the other hand we have

$$\text{C.p.v.} \int\limits_{-\infty}^{+\infty} \frac{x\,dx}{1+x^2} = \lim_{N \to +\infty} \int\limits_{-N}^{N} \frac{x\,dx}{1+x^2} = \frac{1}{2}\ln\frac{1+N^2}{1+N^2} = 0.$$

Hence, the integral in question converges in the sense of the Cauchy principal value. ▶

11.5 Improper Integrals of Unbounded Functions

The necessary condition for the definite (Riemann) integral $\int\limits_a^b f(x)\,dx$ to exist is that the function $f(x)$ be bounded on $[a, b]$. For example, if $f(x)$ is integrable on $[a, b_1]$ where $b_1 < b$ and is unbounded in a neighbourhood of $x = b$ then the definite (Riemann) integral of $f(x)$ does not exist on $[a, b]$. Thus we need some concepts and definitions to generalize the notion of the definite integral to that involving an unbounded integrand.

Let $f(x)$ be integrable on $[a, b - \varepsilon]$ for any whatever small $\varepsilon > 0$ and be unbounded in $(b - \varepsilon, b)$ (Fig. 11.3). To deduce what we mean by writing

$$\int\limits_a^b f(x)\,dx$$

we consider the function in ε ($\varepsilon > 0$)

$$J(\varepsilon) = \int\limits_a^{b-\varepsilon} f(x)\,dx.$$

Definition. If the function $J(\varepsilon)$ has a finite limit L as $\varepsilon \to 0 + 0$ we say that the improper integral $\int\limits_a^b f(x)\,dx$ converges and set

$$\int\limits_a^b f(x)\,dx = \lim_{\varepsilon \to 0+0} \int\limits_a^{b-\varepsilon} f(x)\,dx = L.$$

If the function $J(\varepsilon)$ has no limit as $\varepsilon \to 0 + 0$ the improper integral $\int_a^b f(x)\,dx$ is said to diverge; in this case no numerical value is assigned to this integral.

Example. Investigate the integral $\int_0^1 \frac{dx}{\sqrt{1 - x^2}}$.

◀ The integrand $f(x) = \frac{1}{\sqrt{1 - x^2}}$ is a continuous function on any $[0,\ 1 - \varepsilon]$ where $\varepsilon > 0$. Hence it is integrable. On the other hand

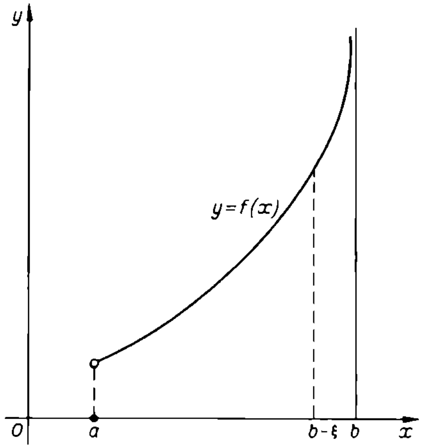


Fig. 11.3

$f(x) \to +\infty$ as $x \to 1 - 0$. Then we have

$$\lim_{\varepsilon\to 0+0} \int_0^{1-\varepsilon} \frac{dx}{\sqrt{1 - x^2}} = \lim_{\varepsilon\to 0+0} (\sin^{-1}(1 - \varepsilon)) = \sin^{-1} 1 = \frac{\pi}{2},$$

so that the improper integral in question converges. Notice that the substitution $x = \sin t$ reduces the original improper integral to $\int_0^{\pi/2} \frac{\cos t}{\cos t}\,dt$, whence we obtain the definite (Riemann) integral $\int_0^{\pi/2} dt = \frac{\pi}{2}$ by putting the integrand equal to 1 at $t = \frac{\pi}{2}$. ▶

Analogously, if a function $f(x)$ is unbounded only in $(a, a + \varepsilon)$ where $\varepsilon > 0$ we define the improper integral $\int_a^b f(x)\,dx$ as

$$\int_a^b f(x)\,dx = \lim_{\varepsilon \to 0+0} \int_{a+\varepsilon}^b f(x)\,dx.$$

This integral is called *convergent* if the finite limit on the right exists; otherwise it is *divergent.*

Example. Investigate the integral $\int_0^1 \frac{dx}{x^\alpha}$.

◀ By the definition we have

$$\int_0^1 \frac{dx}{x^\alpha} = \lim_{\varepsilon \to 0+0} \int_\varepsilon^1 \frac{dx}{x^\alpha}.$$

Since for $\alpha \neq 1$ there holds

$$\int_\varepsilon^1 \frac{dx}{x^\alpha} = \left.\frac{x^{1-\alpha}}{1-\alpha}\right|_\varepsilon^1 = \frac{1}{1-\alpha} - \frac{\varepsilon^{1-\alpha}}{1-\alpha}$$

then

$$\lim_{\varepsilon \to 0+0} \int_\varepsilon^1 \frac{dx}{x^\alpha} = \frac{1}{1-\alpha}$$

provided that $\alpha < 1$.

If $\alpha > 1$ the integral $\int_\varepsilon^1 \frac{dx}{x^\alpha}$ has no finite limit as $\varepsilon \to 0 + 0$.

Let $\alpha = 1$. Then

$$\int_\varepsilon^1 \frac{dx}{x^\alpha} = \int_\varepsilon^1 \frac{dx}{x} = -\ln \varepsilon \to +\infty \text{ as } \varepsilon \to 0 + 0.$$

Thus the integral

$$\int_0^1 \frac{dx}{x^\alpha} \begin{cases} \text{converges} & \text{if } \alpha < 1, \\ \text{diverges} & \text{if } \alpha \geqslant 1. \end{cases}$$

If the function $f(x)$ on $[a, b]$ is unbounded only in some neighbourhood of the point c, where $a < c < b$, then we set

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$$

$$= \lim_{\substack{\varepsilon_1 \to 0+0 \\ \varepsilon_2 \to 0+0}} \left\{ \int_a^{c-\varepsilon_1} f(x)\,dx + \int_{c+\varepsilon_2}^b f(x)\,dx \right\}$$

We leave it to the reader to examine other discontinuities of $f(x)$.

11.6 Improper Integrals of Unbounded Nonnegative Functions. Convergence Tests

Theorem 11.7. *Let $f(x)$ and $\varphi(x)$ be integrable functions on $[a, b - \varepsilon]$ for any whatever small positive ε and be unbounded in $(b - \varepsilon, b)$. Let $0 \leqslant f(x) \leqslant \varphi(x)$ on $[a, b)$. Then*

(1) *if the integral $\int_a^b \varphi(x)\,dx$ converges so does the integral $\int_a^b f(x)\,dx$;*

(2) *if the integral $\int_a^b f(x)\,dx$ diverges so does the integral $\int_a^b \varphi(x)\,dx$.*

◄ Let the integral $\int_a^b \varphi(x)\,dx$ converge so that there exists $\lim_{\varepsilon \to 0+0} \int_a^{b-\varepsilon} \varphi(x)\,dx = L$.

We show that the integral $\int_a^b f(x)\,dx$ converges, i.e., that the function $J(\varepsilon) = \int_a^{b-\varepsilon} f(x)\,dx$ has a finite limit as $\varepsilon \to 0 + 0$. Indeed, since $f(x) \geqslant 0$ on $[a, b)$ then given any $\varepsilon > 0$ ($\varepsilon < b - a$) the function $J(\varepsilon)$ in nonnegative and nondecreasing as ε decreases. Also, if $f(x) \leqslant \varphi(x)$ for all $x \in [a, b)$ then given any $\varepsilon > 0$ there holds

$$\int_a^{b-\varepsilon} f(x)\,dx \leqslant \int_a^{b-\varepsilon} \varphi(x)\,dx.$$

The value of the integral $\int_a^{b-\varepsilon} \varphi(x)\,dx$ does not exceed that of the convergent integral $\int_a^b \varphi(x)\,dx$; hence, given any $\varepsilon > 0$ we have

$$J(\varepsilon) = \int_a^{b-\varepsilon} f(x)\,dx \leqslant \int_a^b \varphi(x)\,dx = L.$$

Therefore the function $J(\varepsilon)$ is nondecreasing as $\varepsilon \to 0 + 0$ and is bounded above. This means that the finite limit of $J(\varepsilon)$ exists as $\varepsilon \to 0 + 0$ and, by virtue of the definition, the improper integral $\int_a^b f(x)\,dx$ converges.

The second part of this theorem can easily be proved by contradiction. ►

Theorem 11.8. *Let* $f(x)$ and $\varphi(x)$ *be positive functions on* $[a, b)$ *and let* $f(x)$ *and* $\varphi(x)$ *be unbounded only in some neighbourhood of the point* $x = b$. *Let there exist*

$$\lim_{x \to b-0} \frac{f(x)}{\varphi(x)} = k > 0.$$

Then both integrals $\int_a^b f(x)\,dx$ *and* $\int_a^b \varphi(x)\,dx$ *either converge or diverge.*

We now turn to the *absolutely convergent improper integrals* involving unbounded functions.

Definition. The integral $\int_a^b f(x)\,dx$ is said to *converge absolutely* if the integral $\int_a^b |f(x)|\,dx$ converges.

As in the case of the improper integrals with infinite limits of integration the following theorem is true.

Theorem 11.9. *If the integral* $\int_a^b |f(x)|\,dx$ *converges so does the integral* $\int_a^b f(x)\,dx$.

Applying this theorem we can easily prove the following convergence test for the improper integrals involving unbounded functions.

Theorem 11.10. *Let* $f(x)$ *be a functon unbounded only in* $(b - \varepsilon, b)$ *where* $\varepsilon > 0$ *is any whatever small number. Let there exist a positive number* $\alpha < 1$ *such that for all* x *sufficiently close to* b, $x < b$, *there holds*

$$|f(x)| \leqslant \frac{M}{(b - x)^{\alpha}},$$

where M *is independent of* x *and* $M > 0$. *Then the integral* $\int_a^b f(x)\,dx$ *converges absolutely.*

Problem. Show that if given any x sufficiently close to b, $x < b$, there holds

$$|f(x)| \geqslant \frac{M}{b - x}, \quad M > 0,$$

then the integral $\int_a^b f(x)\,dx$ does not converge absolutely.

Remark. The substitution $b - x = 1/t$ or $x - a = 1/t$ reduces the improper integrals of unbounded functions to those with infinite limits of integration; so the major results concerning the former integrals can be deduced from the theory of the latter.

11.7 Cauchy Principal Value of the Improper Integral Involving Unbounded Functions

Let $f(x)$ be an integrable function on the closed intervals $[a, c - \varepsilon]$ and $[c + \varepsilon, b]$ where $\varepsilon > 0$, $a < c < b$ and let $f(x)$ be unbounded in some neighbourhood of the point c. As before, we can write

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$$

$$= \lim_{\substack{\varepsilon_1 \to 0+0 \\ \varepsilon_2 \to 0+0}} \left\{ \int_a^{c-\varepsilon_1} f(x)\,dx + \int_{c+\varepsilon_2}^{b} f(x)\,dx \right\}.$$

Notice that the limit on the right must exist as ε_1 and ε_2 tend independently to zero.

Sometimes the improper integral thus defined has no meaning while there exists its Cauchy principal value given by the formula

$$\text{C.p.v.} \int_a^b f(x)\,dx = \lim_{\varepsilon \to 0+0} \left\{ \int_a^{c-\varepsilon} f(x)\,dx + \int_{c+\varepsilon}^{b} f(x)\,dx \right\},$$

where $\varepsilon > 0$ is the same in both integrals on the right. In this case we say that the improper integral converges in the sense of its Cauchy principal value.

Example. Investigate the integral $f(x) = \int_a^b \frac{1}{x - c}$ where $c \in (a, b)$.

◄ We have

$$\int_a^{c-\varepsilon_1} \frac{dx}{x - c} + \int_{c+\varepsilon_2}^{b} \frac{dx}{x - c} = (\ln|x - c|)\Big|_a^{c-\varepsilon_1}$$

$$+ (\ln|x - c|)\Big|_{c+\varepsilon_2}^{b} = \ln\frac{b - c}{c - a} + \ln\frac{\varepsilon_1}{\varepsilon_2}.$$

The limit of the right-hand side as both ε_1 and ε_2 in an arbitrary manner tend to zero does not exist. Put $\varepsilon_1 = \varepsilon_2 = \varepsilon$. Then as $\varepsilon \to 0 + 0$ the limit of the right-hand side exists and is equal to the Cauchy principal value of the original integral so that

$$\text{C.p.v.} \int_a^b \frac{dx}{x-c} = \ln\frac{b-c}{c-a}, \quad a < c < b. \quad \blacktriangleright$$

Problem. Let a function $f(x)$ be defined in the neighbourhood $(-R, R)$ of the point $x = 0$ except probably at $x = 0$ and let $f(x)$ be unbounded as $x \to 0$. It is known that any function $f(x)$ can be represented in a neighbourhood of $x = 0$ by the sum of even and odd functions as

$$f(x) = \frac{f(x) - f(-x)}{2} + \frac{f(x) + f(-x)}{2} = f_1(x) + f_2(x).$$

Show that C.p.v. $\int_{-R}^{R} f(x)\,dx$ exists if so does the integral $\int_0^R f_2(x)\,dx$ where $f_2(x)$ is the even function $\dfrac{f(x) + f(-x)}{2}$.

Exercises

Using the definition of the improper integral determine whether or not the following integrals converge

1. $\displaystyle\int_0^{+\infty} \frac{x\,dx}{x^4+1}$. **2.** $\displaystyle\int_1^{+\infty} \frac{x^2+1}{x^3}\,dx$. **3.** $\displaystyle\int_0^{+\infty} x\mathrm{e}^{-x^2}\,dx$. **4.** $\displaystyle\int_2^{+\infty} \frac{dx}{x\ln x}$.

Determine whether or not the following integrals converge

5. $\displaystyle\int_1^{+\infty} \frac{x\,dx}{x^3+x+1}$. **6.** $\displaystyle\int_{\frac{1}{2}}^{+\infty} \frac{x+1}{x^2+x+5}\,dx$. **7.** $\displaystyle\int_0^{+\infty} \frac{x^2\,dx}{x^4+x^2+1}$.

8. $\displaystyle\int_1^{+\infty} \frac{dx}{\sqrt[3]{x^2+1}}$. **9.** $\displaystyle\int_0^{-\infty} \frac{x\tan^{-1}x}{\sqrt[3]{x^4+1}}\,dx$. **10.** $\displaystyle\int_2^{-\infty} \frac{dx}{x\ln^{\alpha}x}$ (α is a real number). **11.** $\displaystyle\int_1^{+\infty} \frac{\ln\left(1+\frac{1}{x}\right)}{x^{\alpha}}\,dx$.

12. Evaluate the improper integral $\displaystyle\int_0^{+\infty} x^{2n+1}\mathrm{e}^{-x^2}\,dx$ where n is a positive natural number.

13. Show that

$$\text{C.p.v.} \int_{-\infty}^{+\infty} \sin x \, dx = 0 \text{ and C.p.v.} \int_{-\infty}^{+\infty} \frac{dx}{x} = 0.$$

Using the definition of the improper integral determine whether or not the following integrals converge

14. $\displaystyle\int_1^2 \frac{dx}{\sqrt{x-1}}$. **15.** $\displaystyle\int_0^1 x \ln x \, dx$. **16.** $\displaystyle\int_0^1 \frac{dx}{\sqrt[3]{x}}$. **17.** $\displaystyle\int_{-1}^0 \frac{e^{1/x}}{x^3} dx$.

18. $\displaystyle\int_0^1 \frac{e^{1/x}}{x^3} dx$. **19.** $\displaystyle\int_{-1}^1 \frac{dx}{x^4}$.

Investigate the following integrals

20. $\displaystyle\int_0^1 \frac{dx}{\sqrt{1-x^4}}$. **21.** $\displaystyle\int_0^1 \frac{x^3\,dx}{\sqrt[3]{(1-x^2)^5}}$. **22.** $\displaystyle\int_0^1 \frac{dx}{e^{\sqrt{x}}-1}$.

23. $\displaystyle\int_0^1 \frac{dx}{e^x - \cos x}$. **24.** $\displaystyle\int_0^{+\infty} \frac{\tan^{-1} x}{x^\alpha} dx, \ \alpha \geqslant 0$.

25. $\displaystyle\int_0^{+\infty} \frac{x^\alpha}{1+x^\beta} dx$ (α and β are real numbers). **26.** $\displaystyle\int_0^{+\infty} \frac{\sqrt{x}\cos x}{x+1} dx$.

Answer

1. Converges; equal to $\pi/4$. **2.** Diverges. **3.** Converges; equal to 1/2. **4.** Diverges. **5.** Converges. **6.** Diverges. **7.** Converges. **8.** Diverges. **9.** Diverges. **10.** Converges for $\alpha > 1$. **11.** Converges for $\alpha > 0$. **12.** $\dfrac{n!}{2}$. **14.** Converges; equal to 2. **15.** Converges; equal to $-1/4$. **16.** Converges; equal to 3/2. **17.** Converges; equal to $-2/e$. **18.** Diverges. **19.** Diverges. **20.** Converges. **21.** Diverges. **22.** Converges. **23.** Diverges. **24.** *Solution.* The integral in question is a union of the integral with infinite limits of integration and that involving unbounded functions. Indeed, its upper limit is infinite and its integrand $f(x) = \dfrac{\tan^{-1} x}{x^\alpha}$ is not defined at the point $x = 0$ and is unbounded as $x \to 0$ for sufficiently large $\alpha > 0$. We divide the interval of integration into two parts so that the first subinterval would contain the point of discontinuity of $f(x)$ at $x = 0$ and the second one would be appropriate to examine the behaviour of $f(x)$ as $x \to +\infty$. Consider, for example, the semiclosed intervals $(0, 1]$ and

$[1, +\infty)$. Then we have

$$\int_0^{+\infty} \frac{\tan^{-1} x}{x^\alpha}\, dx = \int_0^1 \frac{\tan^{-1} x}{x^\alpha}\, dx + \int_1^{+\infty} \frac{\tan^{-1} x}{x^\alpha}\, dx.$$

Consider the first integral on the right. Recall that $\tan^{-1} x \sim x$ as $x \to 0$; hence, $f(x) = \dfrac{\tan^{-1} x}{x^\alpha}$ is similar to $\dfrac{1}{x^{\alpha-1}}$ near the point $x = 0$. This implies that for the integral $\displaystyle\int_0^1 \frac{\tan^{-1} x}{x^\alpha}\, dx$ to converge it is necessary that $\alpha - 1 < 1$ or $\alpha < 2$. The integrand $f(x) = \dfrac{\tan^{-1} x}{x^\alpha}$ of the second integral on the right behaves similarly to the function $\dfrac{\pi}{2} \cdot \dfrac{1}{x^\alpha}$ as $x \to +\infty$ since $\tan^{-1} x \to \dfrac{\pi}{2}$ as $x \to +\infty$. For this integral to converge it is necessary that $\alpha > 1$. Combining both conditions we infer that both integrals on the right converge only if $1 < \alpha < 2$; this is the condition for the original integral to converge. **25.** Converges if $\alpha > -1$ and $\beta - \alpha > 1$. **26.** Converges conditionally. *Hint:* apply the Dirichlet convergence test.

Chapter 12

Functions of Several Variables

12.1 Basic Notions and Notation

The notion of a function of one variable is somewhat inadequate to model relations and phenomena we come across in the surrounding world. Many real-world functions depend on two or more variables. For example, the volume of the rectangular parallelepiped with edges x, y and z is given by the formula $V = xyz$ where x, y and z are all positive. In this case we have four quantities; the value of V is fully determined by the values of three other variables x, y and z. This is one of the numerous applications of the notion of a function of several variables; the major topic of this chapter is the calculus of functions of several variables.

Let R^n be the n-dimensional Euclidean space and let $M'(x_1', x_2', \ldots, x_n')$ and $M''(x_1'', x_2'', \ldots, x_n'')$ be two points in this space. We let $\varrho(M', M'')$ denote the distance between M' and M'' and define

$$\varrho(M', M'') = \sqrt{\sum_{k=1}^{n} (x_k'' - x_k')^2}\,.$$

If $n = 1$ this formula gives the distance $\varrho(M', M'') = |x_1'' - x_1'|$ between the points $M'(x_1')$ and $M''(x_1'')$ of a line; putting $n = 2$, we obtain the distance $\varrho(M', M'') = \sqrt{(x_1'' - x_1')^2 + (x_2'' - x_2')^2}$ between the points $M'(x_1', x_2')$ and $M''(x_1'', x_2'')$ in a plane.

Definition. Let $M_0(x_1^0, x_2^0, \ldots, x_n^0) \in R^n$ and let ε be a positive real number. The set of all $M \in R^n$ such that $\varrho(M, M_0) < \varepsilon$ is called the *n-dimensional open ball* with centre at M_0 and radius ε.

If $n = 2$ we have $(x - x_0)^2 + (y + y_0)^2 < \varepsilon^2$. This expression defines a circular disk with centre at $M_0(x_0, y_0)$ and radius ε. Notice that the circle bounding this disk is not taken into consideration as shown in Fig. 12.1.

For $n = 3$ we obtain a ball with radius ε and centre at $M_0(x_0, y_0, z_0)$ specified by the equation $(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 < \varepsilon^2$. Again the bounding sphere is out of consideration (Fig. 12.2).

We shall also consider a somewhat different neighbourhood of a point $M_0(x_1^0, x_2^0, \ldots, x_n^0)$ that can be called a *rectangular* neighbourhood. We

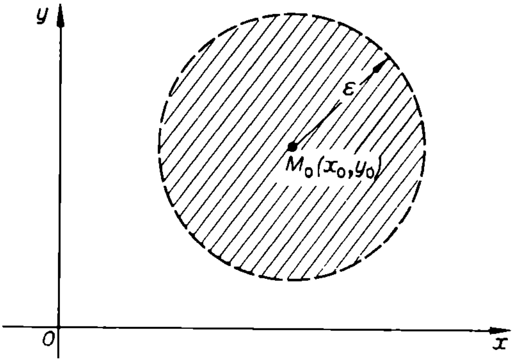


Fig. 12.1

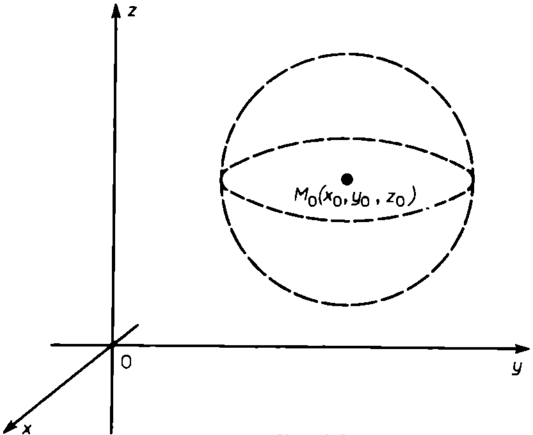


Fig. 12.2

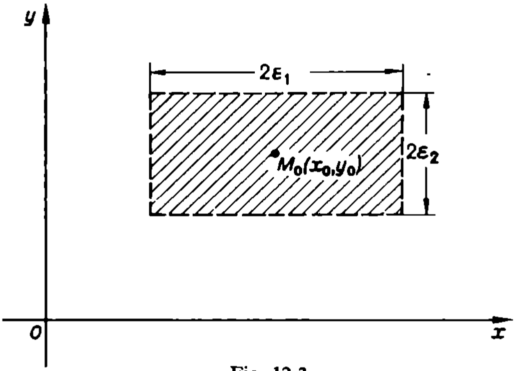


Fig. 12.3

mean here a set of all points $M(x_1, x_2, \ldots, x_n)$ in R^n such that

$$x_i^0 - \varepsilon_i < x_i < x_i^0 + \varepsilon_i,\ \varepsilon_i > 0,\ i = 1, 2, \ldots, n.$$

Putting $n = 1$, we obtain the ε-neighbourhood $x_0 - \varepsilon < x < x_0 + \varepsilon$ of the point x_0 well familiar to us from the calculus of functions of one variable; for $n = 2$ we get the plane figure bounded by the rectangle with sides $2\varepsilon_1$ and $2\varepsilon_2$ (sides are excluded, Fig. 12.3) and for $n = 3$ we have the (open) solid bounded by the parallelepiped with centre at $M_0(x_0, y_0, z_0)$ and edges $2\varepsilon_1$, $2\varepsilon_2$ and $2\varepsilon_3$ (Fig. 12.4).

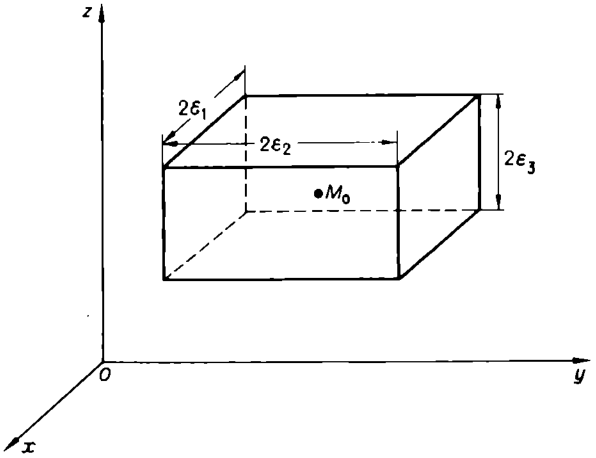


Fig. 12.4

Definitions. Let $E \subset R^n$. The point $M \in E$ is called the *interior point* of the set E if there exists $\varepsilon > 0$ such that E contains M along with its ε-neighbourhood.

The set E is called the *open set* if E contains only interior points. For example, for $n = 2$ any circular disk is an open set.

The point P, $P \in R^n$, is called the *boundary point* of the set $E \subset R^n$ if any neighbourhood of P contains both points belonging to E and those not belonging to E.

The set of all boundary points of E is called the boundary of E; we denote the boundary of E by ∂E.

The union of the set E and its boundary ∂E forms the *closed set* $\overline{E} = E \cup \partial E$. For example, the union of a circular disk and the circle bounding the disk is a closed set.

The set $E \subset R^n$ is called *connected* if given any two points in E a continuous curve (or a polygonal line) joining these points is fully contained only in E (Fig. 12.5*a*). Otherwise the set is called *unconnected* (Fig. 12.5*b*).

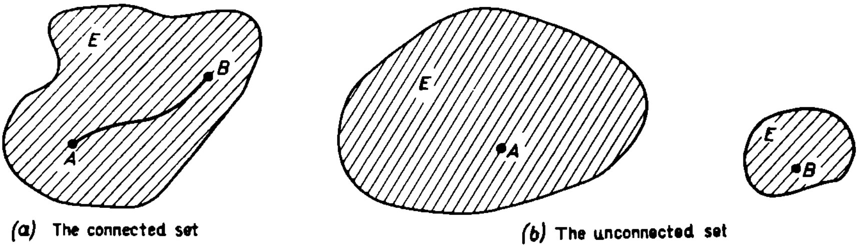
An open connected set is called a *domain.* A domain is said to be *bounded* if there exists a ball that contains this domain.

Any domain that contains a given point M_0 is called a *neighbourhood* of M_0 (in distinction from the ε-neighbourhood of M_0).

Notion of a function of several variables. Suppose that there is known a rule by which every point $M(x_1, x_2, \ldots, x_n)$ of a set E in the n-dimensional Euclidean space R^n is associated with a real number u. Then we say that a function of the point M or a function of n variables $x_1, x_2, \ldots, x_n$ is defined on E and write

$$u = f(M) \text{ or } u = f(x_1, x_2, \ldots, x_n),\ M \in E.$$

The set E is called the *domain of the function f.*



(a) The connected set

(b) The unconnected set

Fig. 12.5

We shall mainly confine our analysis to the functions of two variables $z = f(x, y)$; the major results can easily be generalized to functions of more than two variables.

When a function is specified analytically by a single formula and its domain is not indicated in advance we shall regard the set of all points $M(x_1, x_2, \ldots, x_n)$, where the formula attains a finite real value, as the domain of this function. For example, the domain of the function $z = x + y$ is the xy-plane and the domain of the function $z = \sqrt{1 - x^2 - y^2}$ is the circle $x^2 + y^2 \leqslant 1$.

Let a function $z = f(x, y)$ be defined on a domain E in the xy-plane. Then every point $(x, y) \in E$ corresponds to the point $(x, y, f(x, y))$ in the 3-dimensional space. The set of all points $(x, y, f(x, y))$ where $(x, y) \in E$ is called the *graph* of the function $z = f(x, y)$. For example, the paraboloid of revolution is the graph of the function $z = x^2 + y^2$ (Fig. 12.6).

To investigate the behaviour and visualize the shape of the function $z = f(x, y)$ it is helpful to turn to *level curves.* The level curve is a set of points in the xy-plane where the value of the function is constant so that $z=f(x, y)=c$. It can be constructed by intersecting the surface $z=f(x, y)$

with the plane $z=c$ parallel to the xy-plane as the vertical projection of the line of intersection into the xy-plane. A collection of the level curves $f(x, y) = c_m$, $m = 1, 2, \ldots, k$, where $c_{m+1} - c_m = h = \text{const}$, provides a useful information about the function. It is easily seen that the closer the level curves to each other the higher the rate of change of the function. For the function $z = x^2 + y^2$ its level curves are circles centred at the origin of coordinates (Fig. 12.7, where $h = 1$).

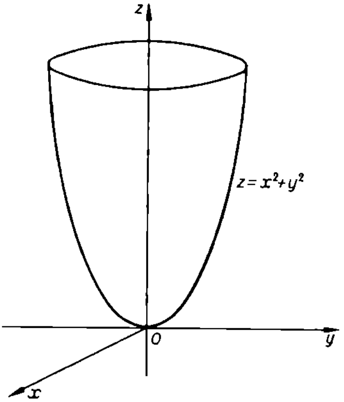


Fig. 12.6

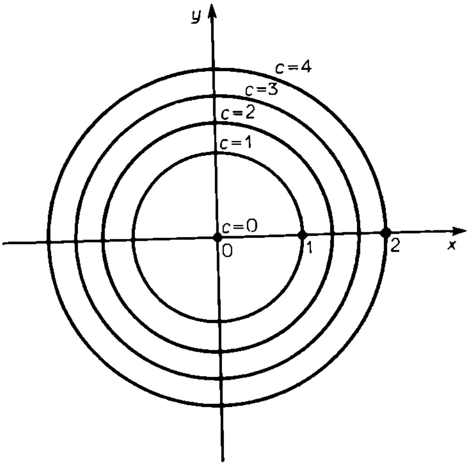


Fig. 12.7

In the case of functions of three variables we can turn to *level surfaces.* The level surface of the function $u = f(x, y, z)$ is the set of points $M(x, y, z)$ in space where the value of $u = f(M)$ is constant. For example, the level surfaces of the function $u = x^2 + y^2 + z^2$ are spheres centred at the origin of coordinates.

12.2 Limits and Continuity

Definition. Let $f(M)$ be a function defined in some neighbourhood Ω of the point $M_0(x_0, y_0)$ except probably at M_0. The number A is called the *limit* of $f(M)$ at the point $M_0(x_0, y_0)$ if given any number $\varepsilon > 0$ there exists a number $\delta > 0$ such that for all points $M(x, y) \in \Omega$ distinct from $M_0(x_0, y_0)$ there holds

$$|f(M) - A| < \varepsilon$$

whenever $0 < \varrho(M, M_0) < \delta$.

In symbols we write

$$A = \lim_{M \to M_0} f(M) \text{ or } A = \lim_{\substack{x \to x_0 \\ y \to y_0}} f(x, y).$$

We assume here that the point M may tend to M_0 in an arbitrary mode (along any arbitrary direction or subject to any arbitrary law) and all the limiting values of $f(M)$ thus obtained must be equal to the number A.

Examples. (1) The function $f(x, y) = x^2 + y^2$ is defined on the xy-plane and $f(0, 0) = 0$. We show that the limit of this function at the point $O(0, 0)$ is equal to zero.

◄ Consider any $\varepsilon > 0$. Then the condition $|f(x, y) - 0| < \varepsilon$ becomes $|x^2 + y^2 - 0| < \varepsilon$ or $|x^2 + y^2| < \varepsilon$. Noting that $\sqrt{x^2 + y^2} = r(M, 0)$ where $M(x, y)$ is a point with the coordinates (x, y), we may write $|x^2 + y^2| < \varepsilon$ as $\varrho^2(M, 0) < \varepsilon$ or $\varrho(M, 0) < \sqrt{\varepsilon}$. Put $\delta = \sqrt{\varepsilon}$. Then given any $M(x, y)$ such that $\varrho(M, 0) < \delta = \sqrt{\varepsilon}$ we have $|x^2 + y^2 - 0| < \varepsilon$ or $|f(x, y) - 0| < \varepsilon$ (Fig. 12.8). By the definition of the limit this means that the number A is the limit of the given function at the point $O(0, 0)$. ►

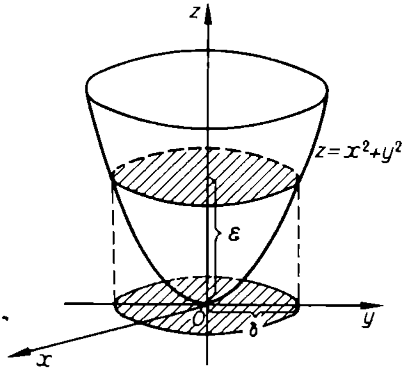


Fig. 12.8

(2) The function $f(x, y) = \dfrac{2xy}{x^2 + y^2}$ is defined everywhere except at the point $O(0, 0)$. We investigate the behaviour of $f(x, y)$ provided that (x, y) tends to zero along the lines $y = kx$, $x \neq 0$.

◄ The lines defined by the equation $y = kx$ pass through the origin of coordinates. We have

$$f(x, kx) = \frac{2x^2k}{(1 + k^2)x^2}, \quad x \neq 0.$$

Whence

$$f(x, kx) \to \frac{2k}{1 + k^2} \text{ as } x \to 0$$

so that for different values of k the limiting values of the function are different. This means that the given function has no limit at the point $O(0, 0)$. ►

(3) The function $f(x, y) = \dfrac{x^2 y}{x^4 + y^2}$ is defined everywhere on the xy-plane except at the origin $O(0, 0)$ of coordinates. Again we investigate the behaviour of this function provided that (x, y) approaches zero along the lines $y = kx$, $x \neq 0$.

◄ We have

$$f(x, kx) = \frac{kx^3}{x^4 + k^2x^2}, \quad x \neq 0$$

so that $f(x, kx) \to 0$ as $x \to 0$. Hence, $f(x, kx)$ has a limit equal to zero for any line $y = kx$, i.e., for any line along which the point tends to the origin of coordinates. If we put $y = x^2$ then $f(x, x^2) = 1/2$, $x \neq 0$. This means that the limit exists when a point tends to the origin by moving along the parabola $y = x^2$. However this limit is equal to 1/2. Thus the given function has no limit at the point $O(0, 0)$. ►

Theorem 12.1. *Let $f(M)$ and $\varphi(M)$ have limits at the point M_0. Then at M_0 there exist the limits of the sum $f(M) + \varphi(M)$, the difference $f(M) - \varphi(M)$, the product $f(M)\varphi(M)$ and the quotient $\dfrac{f(M)}{\varphi(M)}$ (provided that $\lim\limits_{M \to M_0} \varphi(M) \neq 0$) and*

$$\lim_{M \to M_0} (f(M) \pm \varphi(M)) = \lim_{M \to M_0} f(M) \pm \lim_{M \to M_0} \varphi(M),$$

$$\lim_{M \to M_0} (f(M) \cdot \varphi(M)) = \lim_{M \to M_0} f(M) \cdot \lim_{M \to M_0} \varphi(M),$$

$$\lim_{M \to M_0} \frac{f(M)}{\varphi(M)} = \frac{\lim\limits_{M \to M_0} f(M)}{\lim\limits_{M \to M_0} \varphi(M)} \quad (\lim_{M \to M_0} \varphi(M) \neq 0).$$

It is sometimes helpful to use the following definition of the limit which is equivalent to the preceding one.

Definition. Let a function $f(M)$ be defined in some deleted neighbourhood Ω of a point M_0, i.e., in some neighbourhood of M_0 which does not contain M_0. The number A is called the *limit* of the function $f(M)$ at M_0

if given any sequence of points $\{M_n\}$ that converges to M_0 ($M_n \in \Omega$, $M_n \neq M_0$) the corresponding sequence $\{f(M_n)\}$ of the values of $f(M)$ converges to A.

Remark. The above notion of the limit implies that all variables simultaneously tend to their limiting values, i.e., $(x, y) \to (x_0, y_0)$. Sometimes we have to deal with limits obtained by tending variables to their limiting values in succession; there are salient features of functions of several variables. By way of illustration we consider the function $f(x, y) = \dfrac{x^2 - y^2}{x^2 + y^2}$ defined everywhere except at the point $O(0, 0)$. Tending x to zero for $y = \text{const} \neq 0$, we have $\lim\limits_{x \to 0} f(x, y) = -1$. If we put now $x = \text{const} \neq 0$ and tend y to zero then $\lim\limits_{y \to 0} f(x, y) = 1$. Thus we have

$$\lim_{y \to 0} [\lim_{x \to 0} f(x, y)] \neq \lim_{x \to 0} [\lim_{y \to 0} f(x, y)].$$

This means that the limit of $f(x, y)$ depends upon the order in which we tend the variables to their limiting values.

Definition. Let a function $f(M)$ be defined at a point $M_0(x_0, y_0)$ and in some neighbourhood Ω of M_0. The function $f(M)$ is called *continuous* at $M_0(x_0, y_0)$ if $\lim\limits_{M \to M_0} f(M) = f(M_0)$ or, equivalently, if $\lim\limits_{\substack{x \to x_0 \\ y \to y_0}} f(x, y) = f(x_0, y_0)$. We assume here that the point $M(x, y)$ approaches $M_0(x_0, y_0)$ in an arbitrary mode, being always contained in the domain of $f(M)$.

Using the "$\varepsilon\delta$" definition of continuity of a function at a point we say that a function $f(M)$ defined at the point M_0 and in some neighbourhood Ω of M_0 is *continuous* at M_0 if given any $\varepsilon > 0$ there exists $\delta > 0$ such that for all points $M \in \Omega$ there holds $|f(M) - f(M_0)| < \varepsilon$ whenever $\varrho(M, M_0) < \delta$.

We can also give slightly different definition of continuity of a function at a point M_0. We let Δx and Δy denote the increments of the independent variables x and y in moving from the point $M_0(x_0, y_0)$ to the point $M(x, y)$. Let $\Delta z = f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0)$ be the increment in $z = f(x, y)$ correspondng to Δx and Δy. Then the expression

$$\lim_{\substack{x \to x_0 \\ y \to y_0}} f(x, y) = f(x_0, y_0)$$

becomes equivalent to

$$\lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \Delta z = 0,$$

which describes the condition for the function $z = f(x, y)$ to be continuous at the point $M_0(x_0, y_0)$. We assume here again that the increments Δx, Δy can approach zero in an arbitrary manner independently of each other.

The preceding notion refers to the continuity of a function with respect to every and all variables; it implies that if a function $f(x, y)$ is continuous at some point $M_0(x_0, y_0)$ then $f(x, y)$ is continuous at $M_0(x_0, y_0)$ with respect to botn the variable x and the variable y.

However the converse is not always true; if a function $f(x, y)$ is continuous at some point M_0 with respect to both x and y $f(x, y)$ is not necessarily continuous at M_0. To illustrate what we are speaking about we consider the function

$$f(x, y) = \begin{cases} \dfrac{2xy}{x^2 + y^2}, & x^2 + y^2 \neq 0, \\ 0, & x = y = 0. \end{cases}$$

Clearly, we have $f(x, 0) = 0$ for all x, so that $\lim_{x \to 0} f(x, 0) = 0 = f(0, 0)$. Hence, $f(x, 0)$ is continuous with respect to x at $x = 0$. Similarly, the function $f(0, y)$ is continuous with respect to y at $y = 0$ since $f(0, y) = 0$ for all y so that $\lim_{y \to 0} f(0, y) = 0 = f(0, 0)$. However, the function $f(x, y)$ is not continuous at the point $O(0, 0)$. Indeed, let $y = x$. Then

$$\lim_{x = y \to 0} f(x, y) = \lim_{x \to 0} \frac{2x^2}{x^2 + x^2} = 1 \neq f(0, 0).$$

This function fails to exhibit continuity since we have assumed that the variables x and y approach their limiting values in moving along the x- and y-axes only and have left aside infinitely many other directions from which (x, y) may approach the point $O(0, 0)$.

Theorem 12.2. *If both $f(M)$ and $\varphi(M)$ are continuous functions at the point M_0 so are the sum $f(M) + \varphi(M)$, the product $f(M)\varphi(M)$ and the difference $f(M) - \varphi(M)$ and so is the quotient $\dfrac{f(M)}{\varphi(M)}$ provided that $\varphi(M_0) \neq 0$.*

If a function $f(M)$ is continuous at every point of the domain D, $f(M)$ is called *continuous on the domain D.* The point where the function $f(M)$ is not continuous is called the *discontinuity of $f(M)$*. Discontinuities of $f(x, y)$ can either be isolated or fill in some curves. For example, the function $f(x, y) = \dfrac{1}{x^2 + y^2}$ has the only discontinuity $O(0, 0)$ while discontinuities of the function $f(x, y) = \dfrac{1}{x^2 - y^2}$ form the lines $y = x$ and $y = -x$.

Theorem 12.3. *If the function* $f(M)$ *is continuous on the closed bounded domain* D then $f(M)$ *is bounded on* D *and attains its absolute maximum and absolute minimum in* D.

This theorem is a generalization of the respective theorems for functions of one variable (see Sec. 7.3).

12.3 Partial Derivatives and Differentials

Let a function $z = f(x, y)$ be defined on some domain D in the xy-plane and let (x, y) be some interior point of D. Consider the increment Δx in x such that $(x + \Delta x, y) \in D$ (Fig. 12.9).

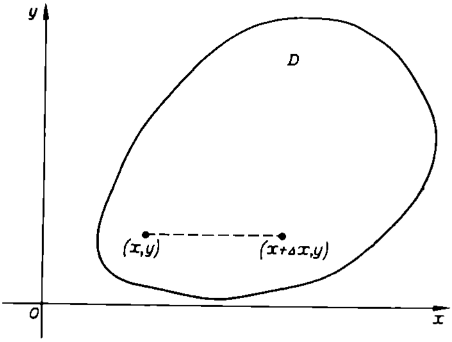


Fig. 12.9

The increment

$$\Delta_x z = f(x + \Delta x, y) - f(x, y)$$

is called the *partial increment* in z caused by the increment Δx in x.

We let $\frac{\Delta_x z}{\Delta x}$ denote the ratio of the partial increment in z to the respective increment in x; clearly, this ratio is a function of Δx.

Definition. If the ratio $\frac{\Delta_x z}{\Delta x}$ has a finite limit as $\Delta x \to 0$ this limit is called the *partial derivative* of the function $z = f(x, y)$ at the given point (x, y) with respect to the independent variable x.

We shall denote the partial derivative of $z = f(x, y)$ with respect to x by writing

$$\frac{\partial z}{\partial x} \text{ or } f_x'(x, y) \text{ or } z_x'(x, y).$$

Therefore we put

$$\frac{\partial z}{\partial x} = \lim_{\Delta x \to 0} \frac{\Delta_x z}{\Delta x}$$

or, equivalently,

$$f'_x(x, y) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x}.$$

Analogously, we have

$$\frac{\partial z}{\partial y} = \lim_{\Delta y \to 0} \frac{\Delta_y z}{\Delta y} = \lim_{\Delta y \to 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y}.$$

Let $u = f(x_1, x_2, \ldots, x_n)$ be a function of n variables. Then

$$\frac{\partial u}{\partial x_k} = \lim_{\Delta x_k \to 0} \frac{f(x_1, x_2, \ldots, x_{k-1}, x_k + \Delta x_k, x_{k+1}, \ldots, x_n) - f(x_1, x_2, \ldots, x_{k-1}, x_k, \ldots, x_n)}{\Delta x_k}.$$

Noting that $\Delta_x z$ and $\Delta_y z$ are computed by regarding x and y, respectively, as constants, we can give the following definitions of partial derivatives.

The partial derivative of a function $z = f(x, y)$ with respect to the variable x is an ordinary derivative with respect to x computed by regarding y as a constant; similarly, the partial derivative of $z = f(x, y)$ with respect to y is an ordinary derivative with respect to y computed by regarding x as a constant. Whence it follows that ordinary and partial derivatives are subject to the same laws of differentiation.

Example. Compute the partial derivatives of the function $z = e^{xy}$.

◄ We have

$$\frac{\partial z}{\partial x} = y e^{xy} \text{ and } \frac{\partial z}{\partial y} = x e^{xy}. \quad ►$$

Remark. If a function $z = f(x, y)$ has partial derivatives with respect to every variable at some point, $z = f(x, y)$ is not necessarily continuous at this point. For example, the function

$$f(x, y) = \begin{cases} \dfrac{xy}{x^2 + y^2}, & x^2 + y^2 \neq 0, \\ 0, & x = y = 0 \end{cases}$$

is not continuous at the point $O(0, 0)$ but has both a partial derivative with respect to x and that with respect to y since $f(x, 0) \equiv 0$ and $f(0, y) \equiv 0$

so that $\left.\dfrac{\partial f}{\partial x}\right|_{(0,0)} = 0$ and $\left.\dfrac{\partial f}{\partial y}\right|_{(0,0)} = 0.$

Geometric interpretation of partial derivatives. Consider a surface S specified by the equation $z = f(x, y)$ in three-dimensional space where $f(x, y)$ is a continuous function having partial derivatives on some domain D. We wish to interpret geometrically partial derivatives of $f(x, y)$ at a point $M_0(x_0, y_0) \in D$, which corresponds to the point $N_0(x_0, y_0, f(x_0, y_0))$ on the surface $z = f(x, y)$.

On computing the partial derivative $\frac{\partial z}{\partial x}$ at M_0 we think of $z = f(x, y)$ as a function of one variable x and treat y as a constant $y = y_0$, i.e., $z = f(x, y_0) = f_1(x)$.

The function $z = f_1(x)$ specifies the curve L obtained by intersecting the surface S with the plane $y = y_0$ (Fig. 12.10). Recalling the geometric interpretation of ordinary derivatives, we can write $f_1'(x_0) = \tan\alpha$ where α is the angle between the x-axis and the tangent to the curve L at the point N_0. Since $f_1'(x_0) = \left(\frac{\partial z}{\partial x}\right)\Big|_{(x_0,\ y_0)}$ then $\left(\frac{\partial z}{\partial x}\right)\Big|_{M_0} = \tan\alpha$.

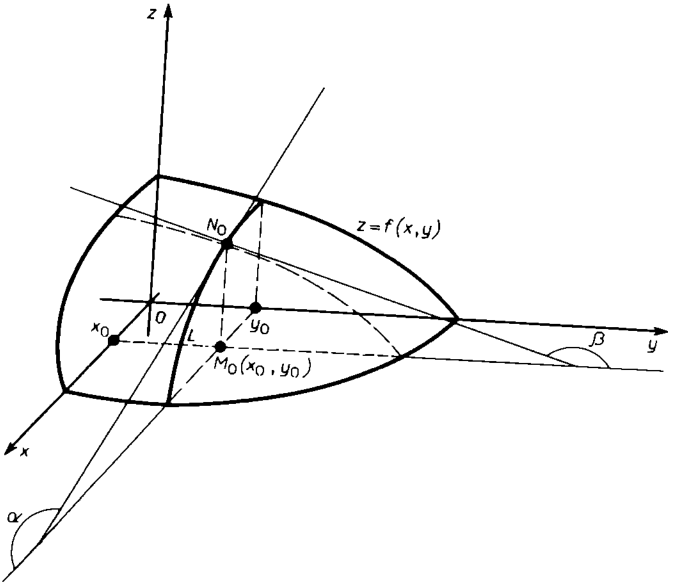


Fig. 12.10

Thus $\left(\frac{\partial z}{\partial z}\right)\Big|_{M_0}$ is the slope of the tangent at N_0 to the curve formed by intersecting the plane $y = y_0$ with the surface $z = f(x, y)$.

Analogously, $\left(\dfrac{\partial z}{\partial y}\right)\Bigg|_{M_0} = \tan\beta$.

Differentiable functions. Let a function $z = f(x, y)$ be defined on some domain D in the xy-plane and let (x, y) be some point in D. We let Δx and Δy denote the increments in x and in y such that $(x + \Delta x, y + \Delta y) \in D$.

Definition. The function $z = f(x, y)$ is said to be *differentiable* at $(x, y) \in D$ if the total increment

$$\Delta z = f(x + \Delta x, y + \Delta y) - f(x, y)$$

corresponding to the increments Δx and Δy admits a representation of the form

$$\Delta z = A\,\Delta x + B\,\Delta y + \alpha(\Delta x, \Delta y)\,\Delta x + \beta(\Delta x, \Delta y)\Delta y, \qquad (*)$$

where A and B are independent of Δx and Δy (though they depend upon x and y in general) and $\alpha(\Delta x, \Delta y)$ and $\beta(\Delta x, \Delta y)$ tend to zero as Δx and Δy approach zero.

If a function $z = f(x, y)$ is differentiable at a point (x, y) the linear (relative to Δx and Δy) part $A\,\Delta x + B\,\Delta y$ of its increment is called the *differential* of $z = f(x, y)$ at the point (x, y).

We denote the differential of a function $z = f(x, y)$ by dz and write

$$dz = A\,\Delta x + B\,\Delta y.$$

Thus, $\Delta z = dz + \alpha\,\Delta x + \beta\,\Delta y$.

Example. Find the differential of the function $z = x^2 + y^2$.

◄ Given any point (x, y) and any increments Δx and Δy there holds

$$\begin{aligned}\Delta z &= (x + \Delta x)^2 + (y + \Delta y)^2 - x^2 - y^2\\ &= 2x\,\Delta x + 2y\,\Delta y + \Delta x\,\Delta x + \Delta y\,\Delta y.\end{aligned}$$

Here we have $A = 2x$, $\beta = 2y$, $\alpha(\Delta x, \Delta y) = \Delta x$ and $\beta(\Delta x, \Delta y) = \Delta y$. Hence α and β tend to zero as $\Delta x \to 0$ and $\Delta y \to 0$.

By the definition of differentiable functions the given function is differentiable at every point in the xy-plane and $dz = 2x\,\Delta x + 2y\,\Delta y$. ►

Notice that we have not excluded the case when either Δx or Δy or both are equal to zero.

We can abbreviate the formula $(*)$ by introducing the distance between the points (x, y) and $(x + \Delta x, y + \Delta y)$ as

$$\varrho = \sqrt{(\Delta x)^2 + (\Delta y)^2}.$$

Then we can write

$$\alpha\,\Delta x + \beta\,\Delta y = \left(\alpha\frac{\Delta x}{\varrho} + \beta\frac{\Delta y}{\varrho}\right)\varrho \quad (\varrho \neq 0)$$

or

$$\alpha\,\Delta x + \beta\,\Delta y = \varepsilon\varrho,$$

where $\varepsilon = \alpha\dfrac{\Delta x}{\varrho} + \beta\dfrac{\Delta y}{\varrho}$ depends upon Δx and Δy and tends to zero as $\Delta x \to 0$ and $\Delta y \to 0$ or, using the abbreviated notation, as $\varrho \to 0$.

Thus formula (*) expressing the condition for the function $z = f(x, y)$ to be differentiable becomes

$$\Delta z = A\,\Delta x + B\,\Delta y + \varepsilon\varrho,$$

where $\varepsilon = \varepsilon(\varrho) \to 0$ as $\varrho \to 0$. For instance, turning to the previous example, we have $\Delta z = 2x\,\Delta x + 2y\,\Delta y + (\Delta x)^2 + (\Delta y)^2 = 2x\,\Delta x + 2y\,\Delta y + \varrho^2$ where $\varepsilon(\varrho) = \varrho$.

Necessary conditions for a function to be differentiable. We shall prove the following important theorems.

Theorem 12.4. *If a function $z = f(x, y)$ is differentiable at some point then $f(x, y)$ is continuous at this point.*

◄ Indeed, if $z = f(x, y)$ is differentiable at a point (x, y) then the increment in z at (x, y) corresponding to the increments Δx and Δy admits a representation of the form

$$\Delta z = A\,\Delta x + B\,\Delta y + \alpha\,\Delta x + \beta\,\Delta y,$$

where A and B are constant at (x, y) and $\alpha \to 0$ and $\beta \to 0$ as $\Delta x \to 0$ and $\Delta y \to 0$, whence $\lim\limits_{\substack{\Delta x \to 0\\ \Delta y \to 0}} \Delta z = 0$. This means that the function $z = f(x, y)$ is continuous at the point (x, y). ►

Theorem 12.5. *If a function $z = f(x, y)$ is differentiable at a point then $f(x, y)$ has partial derivatives $\dfrac{\partial z}{\partial x}$ and $\dfrac{\partial z}{\partial y}$ at this point.*

◄ Indeed, let $z = f(x, y)$ be differentiable at a point (x, y). Then the increment Δz corresponding to the increments Δx and Δy can be written as

$$\Delta z = A\,\Delta x + B\,\Delta y + \alpha(\Delta x, \Delta y)\Delta x + \beta(\Delta x, \Delta y)\Delta y.$$

Put $\Delta x \neq 0$ and $\Delta y = 0$. Then the above formula gives

$$\Delta_x z = A\,\Delta x + \alpha(\Delta x, 0)\Delta x.$$

Whence

$$\frac{\Delta_x z}{\Delta x} = A + \alpha(\Delta x, 0).$$

Since the value of A is independent of Δx and $\alpha(\Delta x, 0) \to 0$ as $\Delta x \to 0$ then

$$A = \lim_{\Delta x \to 0} \frac{\Delta_x z}{\Delta x}.$$

This implies that at the point (x, y) there exists a partial derivative of $z = f(x, y)$ with respect to x and

$$\frac{\partial z}{\partial x} = A.$$

By similar reasoning we infer that at the point (x, y) there exists a partial derivative of $z = f(x, y)$ with respect to y and $\frac{\partial z}{\partial y} = B$. ►

It is worth mentioning here that Theorem 12.5 states that there exist partial derivatives at the point (x, y); however, it tells nothing about continuity and the behaviour of the partial derivatives at (x, y) and in a neighbourhood of the point (x, y).

From Theorem 12.5 it follows that

$$\Delta z = \frac{\partial z}{\partial x}\,\Delta x + \frac{\partial z}{\partial y}\,\Delta y + \alpha\,\Delta x + \beta\,\Delta y.$$

Sufficient conditions for a function to be differentiable. Recall that for a function $y = f(x)$ of one variable to be differentiable at the point x_0 it is necessary and sufficient that $y = f(x)$ have a derivative $f'(x)$ at x_0. However we can not set forth similar necessary and sufficient conditions even for a function $z = f(x, y)$ of two variables; the sufficient conditions exist apart from the necessary conditions discussed above. The former are given by the following theorem.

Theorem 12.6. *Let a function $z = f(x, y)$ have partial derivatives f'_x and f'_y in some neighbourhood of a point (x_0, y_0) and let f'_x and f'_y be continuous at (x_0, y_0). Then $z = f(x, y)$ is differentiable at (x_0, y_0).*

The proof can be found elsewhere in the relevant literature.

Example. Investigate the function $f(x, y) = \sqrt[3]{xy}$.

◄ This function is defined everywhere. By the definition of the partial derivative we have

$$f'_x(0, 0) = \lim_{\Delta x \to 0} \frac{f(\Delta x, 0) - f(0, 0)}{\Delta x} = \lim_{\Delta x \to 0} \frac{\sqrt[3]{\Delta x \cdot 0} - 0}{\Delta x} = 0$$

and

$$f'_y(0, 0) = \lim_{\Delta y \to 0} \frac{f(0, \Delta y) - f(0, 0)}{\Delta y} = \lim_{\Delta y \to 0} \frac{\sqrt[3]{0 \cdot \Delta y} - 0}{\Delta y} = 0.$$

To define whether $f(x, y)$ is differentiable or not at the point $O(0, 0)$ we compute the increment of $f(x, y)$ at $O(0, 0)$; we have

$$\Delta f(0, 0) = \sqrt[3]{\Delta x\,\Delta y} = \varepsilon(\Delta x, \Delta y)\varrho.$$

Since $\varrho = \sqrt{(\Delta x)^2 + (\Delta y)^2}$ then

$$\varepsilon(\Delta x, \Delta y) = \frac{\sqrt[3]{\Delta x\, \Delta y}}{\sqrt{(\Delta x)^2 + (\Delta y)^2}}.$$

For the function $f(x, y)$ to be differentiable at $O(0, 0)$ it is necessary that the function $\varepsilon(\Delta x, \Delta y)$ be an infinitesimal as $\Delta x \to 0$ and $\Delta y \to 0$. Putting $\Delta y = \Delta x > 0$, we can write $\varepsilon(\Delta x, \Delta y)$ as

$$\varepsilon(\Delta x, \Delta y) = \frac{(\Delta x)^{2/3}}{\sqrt{2}\, \Delta x}.$$

It is easy to see that $\varepsilon(\Delta x, \Delta y) \to \infty$ as $\Delta x \to 0$ so that $f(x, y) = \sqrt[3]{xy}$ is not differentiable at $O(0, 0)$ although $f(x, y)$ has partial derivatives f_x' and f_y' at $O(0, 0)$. This result is attributed to the discontinuities of f_x' and f_y' at $O(0, 0)$. ►

Total differential and partial differentials. If a function $z = f(x, y)$ is differentiable then the total differential is given as

$$dz = A\, \Delta x + B\, \Delta y.$$

Noting that $A = \dfrac{\partial z}{\partial x}$ and $B = \dfrac{\partial z}{\partial y}$, we can write dz as

$$dz = \frac{\partial z}{\partial x}\, \Delta x + \frac{\Delta z}{\partial y}\, \Delta y.$$

To extend the notion of the differential we put differentials of independent variables equal to their increments so that

$$dx = \Delta x \text{ and } dy = \Delta y.$$

Then the total differential of $f(x, y)$ can be written in the form

$$dz = \frac{\partial z}{\partial x}\, dx + \frac{\Delta z}{\partial y}\, dy.$$

Example. Compute the differential of the function $z = \ln(x + y^2)$.
◄ We have

$$dz = \frac{1}{x + y^2}\, dx + \frac{2y}{x + y^2}\, dy = \frac{dx + 2y\, dy}{x + y^2}. \quad ►$$

Analogously, if $u = f(x_1, x_2, \ldots, x_n)$ is a differentiable function of n independent variables then

$$du = \sum_{k=1}^{n} \frac{\partial u}{\partial x_k}\, dx_k \quad (dx_k = \Delta x_k).$$

The relation

$$d_x z = f'_x(x, y)\,dx$$

is called the *partial differential of a function* $z = f(x, y)$ *with respect to the variable* x, and

$$d_y z = f'_y(x, y)\,dy$$

is called the *partial differential of* $z = f(x, y)$ *with respect to the variable* y.

Using the above formulas for dz, $d_x z$ and $d_y z$, we can write the differential of $z = f(x, y)$ as

$$dz = d_x z + d_y z,$$

i.e., the total differential of $z = f(x, y)$ is the sum of partial differentials. Notice that in general the total increment Δz in $z = f(x, y)$ is not equal to the sum of the increments in x and y.

Suppose that a function $z = f(x, y)$ is differentiable at the point (x, y) and $dz \neq 0$ at (x, y). Then the total increment

$$\Delta z = \frac{\partial z}{\partial x}\Delta x + \frac{\partial z}{\partial y}\Delta y + \alpha(\Delta x, \Delta y)\Delta x + \beta(\Delta x, \Delta y)\Delta y$$

differs from the linear part

$$dz = \frac{\partial z}{\partial x}\Delta x + \frac{\partial z}{\partial y}\Delta y$$

only by the sum $\alpha\,\Delta x + \beta\,\Delta y$ where Δx and Δy as $\Delta x \to 0$, $\Delta y \to 0$ are infinitesimals of higher orders than the summands in dz; so if $dz \neq 0$ the linear part of the total increment of a differentiable function is called the *principal part* of this increment and the approximation

$$\Delta z \approx dz$$

is widely used. Notice that the smaller the absolute values of the increments in arguments the higher the precision of the approximation.

12.4 Derivatives of Composite Functions

Let a function $z = f(x, y)$ be defined on some domain D in the xy-plane and let both x and y be functions of a variable t so that

$$x = \varphi(t),\ y = \psi(t),\ t_0 < t < t_1.$$

We assume that for any point t in (t_0, t_1) the corresponding point (x, y) is contained in D. Then the substitution $x = \varphi(t)$ and $y = \psi(t)$ reduces $z = f(x, y)$ to the composite function $z = f[\varphi(t), \psi(t)]$ of one variable t.

Theorem 12.7. *If at a point t there exist the derivatives*

$$\frac{dx}{dt} = \varphi'(t) \text{ and } \frac{dy}{dt} = \psi'(t)$$

and a function f(x, y) is differentiable given $x = \varphi(t)$ and $y = \psi(t)$ then the composite function $z = f[\varphi(t), \psi(t)]$ has a derivative $\frac{dz}{dt}$ at t and

$$\frac{dz}{dt} = \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt}.$$

◄ Let Δt be an increment in t which gives rise to the increments Δx and Δy in x and y and, hence, to the increment Δz in the function z. Recall that $z = f(x, y)$ is a differentiable function at (x, y). Then if $(\Delta x)^2 + (\Delta y)^2 \neq 0$ the increment Δz can be written in the form

$$\Delta z = \frac{\partial z}{\partial x}\Delta x + \frac{\partial z}{\partial y}\Delta y + \alpha\,\Delta x + \beta\,\Delta y,$$

where $\alpha(\Delta x, \Delta y)$ and $\beta(\Delta x, \Delta y)$ tend to zero as $\Delta x \to 0$ and $\Delta y \to 0$. If we put $\alpha(0, 0) = 0$ and $\beta(0, 0) = 0$ for $\Delta x = \Delta y = 0$ then $\alpha(\Delta x, \Delta y)$ and $\beta(\Delta x, \Delta y)$ will be continuous as $\Delta x = \Delta y = 0$.

Consider

$$\frac{\Delta z}{\Delta t} = \frac{\partial z}{\partial x}\frac{\Delta x}{\Delta t} + \frac{\partial z}{\partial y}\frac{\Delta y}{\Delta t} + \alpha\frac{\Delta x}{\Delta t} + \beta\frac{\Delta y}{\Delta t}. \qquad (*)$$

Both factors of each summand on the right have limits as $\Delta t \to 0$. Indeed, the partial derivatives $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$ are constant at the given point (x, y). By the hypothesis there exist limits

$$\lim_{\Delta t \to 0}\frac{\Delta x}{\Delta t} = \frac{dx}{dt} = \varphi'(t) \text{ and } \lim_{\Delta t \to 0}\frac{\Delta y}{\Delta t} = \frac{dy}{dt} = \psi'(t).$$

Since $\frac{dx}{dt}$ and $\frac{dy}{dt}$ exist at a point t the functions $x = \varphi(t)$ and $y = \psi(t)$ are continuous at this point. This implies that as $\Delta t \to 0$ both Δx and Δy tend to zero and, hence, so do $\alpha(\Delta x, \Delta y)$ and $\beta(\Delta x, \Delta y)$.

Thus the right-hand side of (*) has a limit as $\Delta t \to 0$ equal to $\frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt}$. This means that the left-hand side of (*) also has a limit as $\Delta t \to 0$ so that there exists $\lim\limits_{\Delta t \to 0}\frac{\Delta z}{\Delta t} = \frac{dz}{dt}$. Evaluating the limits in (*) as $\Delta t \to 0$, we obtain

$$\frac{dz}{dt} = \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt}. \quad ► \qquad (**)$$

Example. Compute the derivative of the function $z = x^2 + y^2$ where $x = \sin t$ and $y = t^3$.

◀ Using the above formula, we get

$$\frac{dz}{dt} = 2x \cos t + 2y\, 3t^2 = 2 \sin t \cos t + 6t^5 = \sin 2t + 6t^5. \quad ▶$$

Let us consider a function $z = f(x, y)$ where $y = \psi(x)$, z being a composite function of x so that $z = f(x, \psi(x))$. Then

$$\frac{dz}{dx} = \frac{\partial z}{\partial x} + \frac{\partial z}{\partial y}\frac{dz}{dx},$$

where $\frac{\partial z}{\partial x}$ is a partial derivative of $z = f(x, y)$ with respect to x computed by regarding the variable y as a constant. The derivative $\frac{dz}{dx}$ is the *total* derivative of $z = f(x, y)$ with respect to the independent variable x computed by regarding y as a function of x, i.e., $y = \psi(x)$ so that the relationship between z and x is taken into account.

Example. Compute $\frac{\partial z}{\partial x}$ and $\frac{dz}{dx}$ provided that $z = \tan^{-1}\frac{y}{x}$ and $y = x^2$.

$$◀ \quad \frac{\partial z}{\partial x} = \frac{\partial}{\partial x}\left(\tan^{-1}\frac{y}{x}\right) = -\frac{y}{x^2 + y^2}, \quad \frac{\partial z}{\partial y} = \frac{x}{x^2 + y^2}$$

and $$\frac{dz}{dx} = \frac{\partial z}{\partial x} + \frac{\partial z}{\partial y}\frac{dy}{dx} = -\frac{y}{x^2 + y^2} + \frac{x}{x^2 + y^2}2x = \frac{1}{1 + x^2}. \quad ▶$$

Now we draw our attention to differentiation of a composite function of several variables. Consider the function $z = f(x, y)$ where $x = \varphi(\xi, \eta)$ and $y = \psi(\xi, \eta)$ so that $z = z(\xi, \eta) = f(\varphi(\xi\, \eta), \psi(\xi, \eta))$. Let continuous partial derivatives $\frac{\partial x}{\partial \xi}, \frac{\partial x}{\partial \eta}, \frac{\partial y}{\partial \xi}$ and $\frac{\partial y}{\partial \eta}$ exist at a point (ξ, η) and let $f(x, y)$ be differentiable at the corresponding point (x, y) where $x = \varphi(\xi, \eta)$ and $y = \psi(\xi, \eta)$.

We show that the composite function $z = z(\xi, \eta)$ has the derivatives $\frac{\partial z}{\partial \xi}$ and $\frac{\partial z}{\partial \eta}$ at the point (ξ, η) and derive the respective formulas. Notice that this case is almost similar to that considered before. Indeed, on differentiating z with respect to ξ the variable η is regarded as a constant; this implies that x and y become functions of one variable ξ, i.e., $x = \varphi(\xi, c)$ and $y = \psi(\xi, c)$ so that formula (**) is fully applicable here.

Using (**), we obtain

$$\frac{\partial z}{\partial \xi} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial \xi} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial \xi}.$$

Analogously

$$\frac{\partial z}{\partial \eta} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial \eta} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial \eta}.$$

Example. Compute the partial derivatives $\frac{\partial z}{\partial \xi}$ and $\frac{\partial z}{\partial \eta}$ of the function $z = x^2y - xy^2$ where $x = \xi\eta$ and $y = \frac{\xi}{\eta}$.

◄ $$\frac{\partial z}{\partial \xi} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial \xi} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial \xi} = (2xy - y^2)\eta + (x^2 - 2xy)\frac{1}{\eta}$$

$$= \left(2\xi\eta\frac{\xi}{\eta} - \frac{\xi^2}{\eta^2}\right)\eta + \left(\xi^2\eta^2 - 2\xi\eta\frac{\xi}{\eta}\right)\frac{1}{\eta} = 3\xi^2\left(\eta - \frac{1}{\eta}\right)$$

and

$$\frac{\partial z}{\partial \eta} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial \eta} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial \eta} = (2xy - y^2)\xi + (x^2 - 2xy)\left(-\frac{\xi}{\eta^2}\right)$$

$$= \left(2\xi\eta\frac{\xi}{\eta} - \frac{\xi^2}{\eta^2}\right)\xi + \left(\xi^2\eta^2 - 2\xi\eta\frac{\xi}{\eta}\right)\left(-\frac{\xi}{\eta^2}\right) = \xi^3\left(1 + \frac{1}{\eta^2}\right).$$ ►

If a composite function $u = f(x, y, z)$ is given by $x = x(\xi, \eta)$, $y = y(\xi, \eta)$ and $z = z(\xi, \eta)$ so that $u = u(\xi, \eta)$ then

$$\frac{\partial u}{\partial \xi} = \frac{\partial u}{\partial x}\frac{\partial x}{\partial \xi} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial \xi} + \frac{\partial u}{\partial z}\frac{\partial z}{\partial \xi}$$

and

$$\frac{\partial u}{\partial \eta} = \frac{\partial u}{\partial x}\frac{\partial x}{\partial \eta} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial \eta} + \frac{\partial u}{\partial z}\frac{\partial z}{\partial \eta}.$$

In particular, if $u = f(x, y, z)$ where $z = z(x, y)$ then

$$\frac{\partial u}{\partial x} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial x} \text{ and } \frac{\partial u}{\partial y} = \frac{\partial f}{\partial y} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial y}.$$

Here $\frac{\partial u}{\partial x}$ is a total partial derivative of u with respect to the independent variable x allowing for the relationship between u and x, $\frac{\partial f}{\partial x}$ is a partial derivative of $u = f(x, y, z)$ with respect to x computed by regarding the variables y and z as constants. The meanings of $\frac{\partial u}{\partial y}$ and $\frac{\partial f}{\partial y}$ are similar to those of $\frac{\partial u}{\partial x}$ and $\frac{\partial f}{\partial y}$, respectively.

Differentials of composite functions. If $z = f(x, y)$ is a differentiable function of independent variables x and y then the total differential dz of z is given by

$$dz = \frac{\partial z}{\partial x}\, dx + \frac{\partial z}{\partial y}\, dy, \tag{*}$$

where $dx = \Delta x$ and $dy = \Delta y$.

Now we suppose that $z = f(x, y)$ is a composite function. Assume that $x = \varphi(\xi, \eta)$ and $y = \psi(\xi, \eta)$ and that these functions have continuous partial derivatives with respect to ξ and η at a point (ξ, η). Let there exist continuous partial derivatives $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$ at the point (x, y) corresponding to (ξ, η) so that $z = f(x, y)$ is differentiable at (x, y). Then the function $z = f[\varphi(\xi, \eta), \psi(\xi, \eta)]$ has at the point (ξ, η) the derivatives

$$\frac{\partial z}{\partial \xi} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial \xi} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial \xi}$$

and

$$\frac{\partial z}{\partial \eta} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial \eta} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial \eta}.$$

It is easy to see that $\frac{\partial z}{\partial \xi}$ and $\frac{\partial z}{\partial \eta}$ are continuous at the the point (ξ, η). Hence, the function $z = f[\varphi(\xi, \eta), \psi(\xi, \eta)]$ is differentiable at (ξ, η) and by virtue of the formula for the total differential of independent variables η and ξ we have

$$dz = \frac{\partial z}{\partial \xi}\, d\xi + \frac{\partial z}{\partial \eta}\, d\eta.$$

Substituting the expressions for $\frac{\partial z}{\partial \xi}$ and $\frac{\partial z}{\partial \eta}$ into the above formula, we get

$$dz = \left(\frac{\partial z}{\partial x}\frac{\partial x}{\partial \xi} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial \xi}\right)d\xi + \left(\frac{\partial z}{\partial x}\frac{\partial x}{\partial \eta} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial \eta}\right)d\eta$$

or

$$dz = \frac{\partial z}{\partial x}\left(\frac{\partial x}{\partial \xi}\, d\xi + \frac{\partial x}{\partial \eta}\, d\eta\right) + \frac{\partial z}{\partial y}\left(\frac{\partial y}{\partial \xi}\, d\xi + \frac{\partial y}{\partial \eta}\, d\eta\right).$$

By the hypothesis the functions $x = \varphi(\xi, \eta)$ and $y = \psi(\xi, \eta)$ have continuous partial derivatives at the point (ξ, η). Then they are differentiable at (ξ, η) and

$$dx = \frac{\partial x}{\partial \xi}\, d\xi + \frac{\partial x}{\partial \eta}\, d\eta \text{ and } dy = \frac{\partial y}{\partial \xi}\, d\xi + \frac{\partial y}{\partial \eta}\, d\eta.$$

From the above formulas for dz and dx it easily follows that

$$dz = \frac{\partial z}{\partial x}\,dx + \frac{\partial z}{\partial y}\,dy.$$

Comparing the above formula with (*) we infer that the total differential of the function $z = f(x, y)$ is expressed by formulas of the same form when the variables x and y are independent and when x and y are functions of some other variables. Thus the form of the expression for the total differential of a function of two and more variables remains the same (invariant) in both these cases.

Remark. Based upon the previous results it is easy to verify that the differentiation formulas

$$d(x \pm y) = dx \pm dy,$$

$$d(xy) = x\,dy + y\,dx,$$

$$d\left(\frac{x}{y}\right) = \frac{y\,dx - x\,dy}{y^2}.$$

remain valid for x and y being differentiable functions of any finite number of variables, i.e., for $x = \varphi(\xi, \eta, \zeta, \ldots)$ and $y = \psi(\xi, \eta, \zeta, \ldots)$.

12.5 Implicit Functions

Consider the equation $F(x, y) = 0$ where $F(x, y)$ is a function of two variables defined on the domain G in the xy-plane. We suppose that for every x in some interval $(x_0 - h_0, x_0 + h_0)$ there exists the only value of y which satisfies the equation $F(x, y) = 0$; clearly in this case the function $y = y(x)$ such that the identity $F(x, y(x)) = 0$ is fulfilled for all x in $(x_0 - h_0, x_0 + h)$ is fully determined. Then we say that the equation $F(x, y) = 0$ specifies the quantity y as an *implicit function* of x.

In other words, the function $y = y(x)$ specified by the equation $F(x, y) = 0$ involving y is called the implicit function. To express the relationship between y and x explicitly we have to solve the original equation with respect to y.

Examples. **(1)** The equation $y - x = 0$ specifies y as a single-valued function $y = x$ of x at every point of the x-axis.

(2) The equation $y - x - \varepsilon \sin y = 0$, $0 < \varepsilon < 1$, specifies y as a single-valued function of x.

◄ To verify this we investigate the original equation $y - x - \varepsilon \sin y = 0$. Notice that this equation is satisfied if $x = 0$ and $y = 0$. We regard x as a parameter and consider the functions $z = y$ and $z = x + \varepsilon \sin y$. It is clear that if for a given x_0 there exists the only value

y_0 such that the point (x_0, y_0) satisfies the original equation, then the line $z = y$ and the curve $z = x_0 + \varepsilon \sin y$ intersect at a unique point and vice versa. Let us draw the graphs of $z = y$ and $z = x_0 + \varepsilon \sin y$ in the zy-plane (Fig. 12.11).

The graph of $z = x + \varepsilon \sin y$ where x is thought of as a parameter is obtained by translating the graph of the curve $z = \varepsilon \sin y$ along the z-axis. It is easy to see that the graphs of $z = y$ and $z = x + \varepsilon \sin y$ meet at the only point whose ordinate y is a function of x implicitly specified by the equation $y - x - \varepsilon \sin y = 0$, $0 < \varepsilon < 1$. This relationship can not be expressed in elementary functions. ►

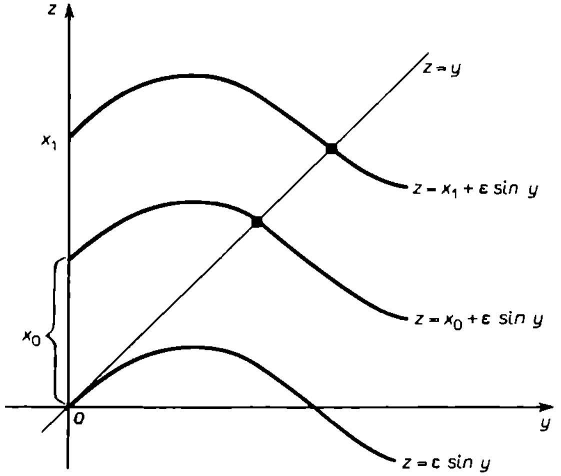

z
$z = y$
x_1
$z = x_1 + \varepsilon \sin y$
x_0
$z = x_0 + \varepsilon \sin y$
O
y
$z = \varepsilon \sin y$

Fig. 12.11

(3) The equation $x^2 + y^2 + 1 = 0$ does not specify y as a real-valued function of x whatever is a real value of x.

Analogously we can introduce the notion of an implicit function of several variables.

The following theorem specifies the sufficient conditions for the equation $F(x, y) = 0$ to be solvable for y in some neighbourhood of a given point x_0.

Theorem 12.8. *Let there be given an equation* $F(x, y) = 0$ *and let the following conditions be satisfied:*

(i) *the function* $F(x, y)$ *is defined and continuous in the rectangle* $D\{x_0 - \delta_1 < x < x_0 + \delta_1,\ y_0 - \delta_2 < y < y_0 + \delta_2\}$, $\delta_1 > 0$, $\delta_2 > 0$, *with centre at* (x_0, y_0);

(ii) *the function* $F(x, y)$ *vanishes at the point* (x_0, y_0) *so that* $F(x_0, y_0) = 0$;

(iii) *there exist continuous partial derivatives* $\frac{\partial F}{\partial x}$ *and* $\frac{\partial F}{\partial y}$ *in* D;

(iv) $\frac{\partial F(x_0, y_0)}{\partial y} \neq 0$.

Then given any sufficiently small positive number ε *there exists a neighbourhood* $x_0 - \delta_0 < x_0 < x_0 + \delta_0$ *of* x_0, $\delta_0 > 0$, *where there exists a uniquely defined* [*] *continuous function* $y = f(x)$ (Fig. 12.12) *such that* $y_0 = f(x_0)$, $|y - y_0| < \varepsilon$ *and the equation* $F(x, f(x)) = 0$ *becomes an identity, i.e.,*

$$F(x, f(x)) \equiv 0$$

for all x *in the given neighbourhood of* x_0.

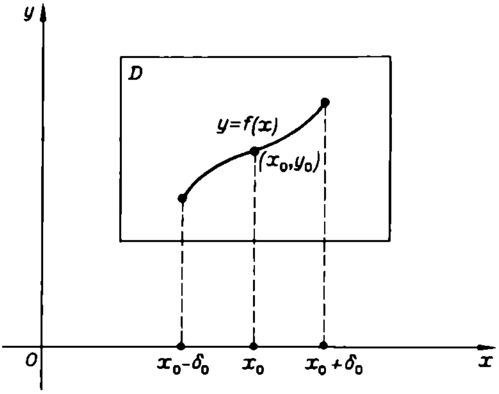


Fig. 12.12

The function $y = f(x)$ *is continuously differentiable in the neighbourhood of* x_0 *and*

$$\frac{dy}{dx} = -\frac{\frac{\partial F}{\partial x}}{\frac{\partial F}{\partial y}} \quad \left(\frac{\partial F}{\partial y} \neq 0\right). \qquad (*)$$

---

[*] We regard $y = f(x)$ as "uniquely defined" in the sense that the coordinates of any point which lies on the curve $F(x, y) = 0$ and is contained in the neighbourhood $\Omega = \{x_0 - \delta_0 < x < x_0 + \delta_0, y_0 - \varepsilon < y < y_0 + \varepsilon\}$ of the point (x_0, y_0) are related by the equation $y = f(x)$.

◀ We deduce (∗) for an arbitrary implicit function assuming that the derivative $\frac{dy}{dx}$ exists.

Let $y = f(x)$ be a continuous implicit function given by the equation $F(x, y) = 0$. Then for any x in $(x_0 - \delta_0, x_0 + \delta_0)$ there holds the identity $F(x, f(x)) \equiv 0$; hence $\frac{dF(x, f(x))}{dx} = 0$ for any point in this neighbourhood.

By the chain rule of differentiating a composite function we have

$$\frac{dF(x, y(x))}{dx} = \frac{\partial F}{\partial x} + \frac{\partial F}{\partial y}\frac{dy}{dx}.$$

Therefore for $y = f(x)$ we get

$$\frac{\partial F}{\partial x} + \frac{\partial F}{\partial y}\frac{dy}{dx} = 0.$$

Whence

$$\frac{dy}{dx} = -\frac{\partial F/\partial x}{\partial F/\partial y}. \quad ▶$$

Example. Compute $\frac{dy}{dx}$ of the function $y = y(x)$ defined by the equation $x^2 + y^2 = R^2$.

◀ Here we have $F(x, y) = x^2 + y^2 - R^2$ and $\frac{\partial F}{\partial x} = 2x$ and $\frac{\partial F}{\partial y} = 2y$. Using (∗), we obtain $\frac{dy}{dx} = -\frac{x}{y}$ $(y \neq 0)$. ▶

Remark. Theorem 12.8 specifies the sufficient conditions for an implicit function whose graph passes through a given point (x_0, y_0) to exist. However it tells nothing about the necessary conditions. Indeed, consider the equation $(y - x)^2 = 0$. Here the function $F(x, y) = (y - x)^2$ has continuous partial derivatives F'_x and F'_y and $F'_y = 2(y - x)$ is equal to zero at the point $O(0, 0)$; nevertheless the equation $F(x, y) = (y - x)^2$ has the unique solution $y = x$ vanishing at $x = 0$.

Problem. Consider the equation $y^2 = x^2$. Let $y = y(x)$, $-\infty < x < +\infty$, be a single-valued function satisfying $y^2 = x^2$.

Find out how many (1) single-valued functions of the form $y = y(x)$, $-\infty < x < +\infty$; (2) single-valued continuous functions; (3) single-valued differentiable functions; (4) single-valued continuous functions $y = y(x)$, $1 - \delta < x < 1 + \delta$, if $y(1) = 1$ and $\delta > 0$ is sufficiently small, satisfy the equation $y^2 = x^2$.

Similarly to Theorem 12.8 we can formulate a theorem for an implicit function $z = z(x, y)$ of two variables specified by the equation $F(x, y, z) = 0$.

Theorem 12.9. *Let there be given an equation* $F(x, y, z) = 0$ *and let the following conditions be satisfied:*

(i) *the function* $F(x, y, z)$ *is defined and continuous on the domain* D *such that*

$$D = \begin{cases} x_0 - \delta_1 < x < x_0 + \delta_1, \\ y_0 - \delta_2 < y < y_0 + \delta_2, \\ z_0 - \delta_3 < z < z_0 + \delta_3, \end{cases}$$

where $\delta_1 > 0$, $\delta_2 > 0$ *and* $\delta_3 > 0$;

(ii) $F(x_0, y_0, z_0) = 0$;

(iii) *there exist continuous partial derivatives* F'_x, F'_y *and* F'_z *in* D;

(iv) $F'_z(x_0, y_0, z_0) \neq 0$.

Then given any sufficiently small $\varepsilon > 0$ *there exists a neighbourhood* Ω *of the point* (x_0, y_0) *where there exists a uniquely defined continuous function* $z = f(x, y)$ *such that* $z = f(x, y)$ *attains the value* z_0 *at* $x = x_0$ *and* $y = y_0$, $|z - z_0| < \varepsilon$ *and the equation* $F(x, y, z) = 0$ *becomes an identity, i.e.,* $F(x, y, f(x, y)) \equiv 0$ *for all* (x, y) *in* Ω. *The function* $z = f(x, y)$ *has continuous partial derivatives* f'_x *and* f'_y *in* Ω.

Let us deduce formulas for f'_x and f'_y. Suppose that the equation $F(x, y, z) = 0$ specifies z as a single-valued differentiable function $z = f(x, y)$ of independent variables x and y. If we substitute $f(x, y)$ for z then $F(x, y, z) = 0$ becomes the identity

$$F(x, y, f(x, y)) \equiv 0 \quad (x, y) \in \Omega.$$

Hence, the partial derivatives of $F(x, y, z)$ where $z = f(x, y)$ with respect to x and y must also be equal to zero. On differentiating, we get

$$\frac{\partial F}{\partial x} + \frac{\partial F}{\partial z}\frac{\partial z}{\partial x} = 0 \text{ and } \frac{\partial F}{\partial y} + \frac{\partial F}{\partial z}\frac{\partial z}{\partial y} = 0.$$

Whence

$$\frac{\partial z}{\partial x} = -\frac{\partial F/\partial x}{\partial F/\partial z} \text{ and } \frac{\partial z}{\partial y} = -\frac{\partial F/\partial y}{\partial F/\partial z}\left(\frac{\partial F}{\partial z} \neq 0\right).$$

These are the formulas for partial derivatives of an implicit function of two independent variables.

Example. Compute the partial derivatives of a function $z(x, y)$ given by the equation $F(x, y, z) \equiv x^2 + y^2 + z^2 - R^2 = 0$.

◄ We have $\frac{\partial F}{\partial x} = 2x$, $\frac{\partial F}{\partial y} = 2y$ and $\frac{\partial F}{\partial z} = 2z$; whence $\frac{\partial z}{\partial x} = -\frac{x}{z}$ and $\frac{\partial z}{\partial y} = -\frac{y}{z}$ $(z \neq 0)$. ►

To verify the results we can find an explicit formula for the function and apply standard formulas for partial derivatives directly.

12.6 Tangent Planes and Normal Lines to a Surface

Let S be a surface given by the equation $F(x, y, z) = 0$.

Definition. The point $M(x, y, z)$ on a surface S is called the *regular* (*nonsingular*) *point* of S if all the three partial derivatives $\frac{\partial F}{\partial x}$, $\frac{\partial F}{\partial y}$ and $\frac{\partial F}{\partial z}$ at M exist and are continuous and at least one of them is distinct from zero.

The point $M(x, y, z)$ is called the *singular point* of S if all the three partial derivatives $\frac{\partial F}{\partial x}$, $\frac{\partial F}{\partial y}$ and $\frac{\partial F}{\partial z}$ vanish at M or at least one of them does not exist at M.

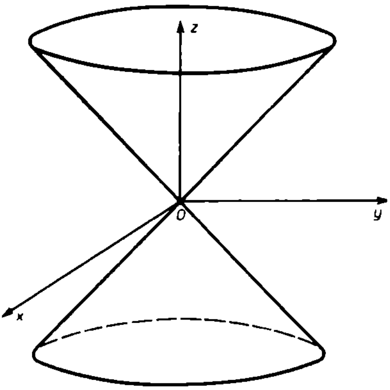


Fig. 12.13

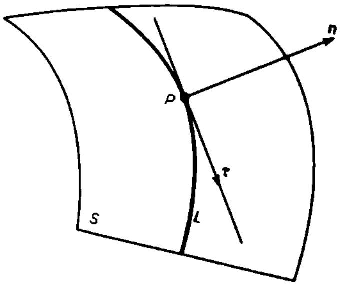


Fig. 12.14

Example. Consider a circular cone given by the equation $x^2 + y^2 - z^2 = 0$ (Fig. 12.13).

◄ Here we have $F(x, y, z) = x^2 + y^2 - z^2$ so that $\frac{\partial F}{\partial x} = 2x$, $\frac{\partial F}{\partial y} = 2y$ and $\frac{\partial F}{\partial z} = -2z$. The only singular point is the origin of coordinates $O(0, 0, 0)$ where all the three partial derivatives vanish. ►

Let L be a space curve given by the parametric equations

$$\begin{cases} x = \varphi(t), \\ y = \psi(t), \quad \alpha < t < \beta. \\ z = \omega(t), \end{cases}$$

Suppose that $\varphi(t)$, $\psi(t)$ and $\omega(t)$ have continuous derivatives $\varphi'(t)$, $\psi'(t)$ and $\omega'(t)$ at every t such that $\alpha < t < \beta$. We leave aside singular points of L where $\varphi'^2(t), + \psi'^2(t) + \omega'^2(t) = 0$ and consider a regular point $M_0(x_0, y_0, z_0)$ on L specified by the value t_0 of t, $t_0 \in (\alpha, \beta)$. Then the vector $\tau = x'(t_0)\mathbf{i} + y'(t_0)\mathbf{j} + z'(t_0)\mathbf{k}$ lies on the tangent to the curve L at the point $M_0(x_0, y_0, z_0)$.

Now we choose a regular point P of a surface S and draw through P a curve L lying in S. Let the curve be given, as before, by the parametric equations

$$\begin{cases} x = \varphi(t), \\ y = \psi(t), & \alpha < t < \beta \\ z = \omega(t), \end{cases}$$

and let $\varphi(t)$, $\psi(t)$ and $\omega(t)$ have continuous derivatives that nowhere in $\alpha(\beta)$ vanish simultaneously.

By the definition the tangent to L at P is called the *tangent* to the surface S at P.

If the above parametric equations are substituted into the equation $F(x, y, z) = 0$ of the surface S the latter becomes an identity with respect to t so that $F(\varphi(t), \psi(t), \omega(t)) \equiv 0$ since the curve L lies in the surface S.

Differentiating this identity as a composite function of t, we get

$$\frac{\partial F}{\partial x}\frac{dx}{dt} + \frac{\partial F}{\partial y}\frac{dy}{dt} + \frac{\partial F}{\partial z}\frac{dz}{dt} = 0. \qquad (*)$$

The expression on the left is a scalar product of the vectors

$$\mathbf{n} = \frac{\partial F}{\partial x}\mathbf{i} + \frac{\partial F}{\partial y}\mathbf{j} + \frac{\partial F}{\partial z}\mathbf{k} \text{ and } \tau = \frac{dx}{dt}\mathbf{i} + \frac{dy}{dt}\mathbf{j} + \frac{dz}{dt}k.$$

The vector τ is a tangent vector to the curve L at the point P. The vector $\mathbf{n}$ being independent of the shape of the curve passing through the point P depends only on the coordinates of P and on the shape of the function $F(x, y, z)$.

Since P is a regular point the length of $\mathbf{n}$

$$|\mathbf{n}| = \sqrt{\left(\frac{\partial F}{\partial x}\right)^2 + \left(\frac{\partial F}{\partial y}\right)^2 + \left(\frac{\partial F}{\partial z}\right)^2}$$

is distinct from zero.

The identity $(*)$ implies that $(\mathbf{n}, \tau) = 0$. This means that the vector τ tangent to the curve L at the point P is perpendicular to the vector $\mathbf{n}$ at P (Fig. 12.14). The same reasoning is fully applicable to any curve lying in the surface S and passing through the point P. Hence, any tangent to the surface S at the point P is perpendicular to the vector $\mathbf{n}$ and thus all such tangents lie in the same plane perpendicular to the vector $\mathbf{n}$.

Definition. The plane formed by all tangents to a surface S through a given regular point $P \in S$ is called the *tangent plane* to the surface at P.

The vector $\mathbf{n}\Big|_P = \left\{\dfrac{\partial F}{\partial x}\Big|_P, \dfrac{\partial F}{\partial y}\Big|_P, \dfrac{\partial F}{\partial z}\Big|_P\right\}$ is a *normal vector* of the tangent plane $F(x, y, z) = 0$ at the point P. Whence we get the equation of the tangent plane to the surface $F(x, y, z) = 0$ at the regular point $P_0(x_0, y_0, z_0)$ as

$$\left(\frac{\partial F}{\partial x}\right)\Bigg|_{(x_0,y_0,z_0)}(x-x_0)+\left(\frac{\partial F}{\partial y}\right)\Bigg|_{(x_0,y_0,z_0)}(y-y_0)+\left(\frac{\partial F}{\partial z}\right)\Bigg|_{(x_0,y_0,z_0)}(z-z_0)=0.$$

If the surface S is given by the equation $z = f(x, y)$ we can write it as $F \equiv f(x, y) - z = 0$; whence $\dfrac{\partial F}{\partial x} = \dfrac{\partial f}{\partial x}$, $\dfrac{\partial F}{\partial y} = \dfrac{\partial f}{\partial y}$ and $\dfrac{\partial F}{\partial y} = -1$. Then the equation $z_0 = f(x_0, y_0)$ of the tangent plane at the point $P_0(x_0, y_0, z_0)$ becomes

$$z - z_0 = \left(\frac{\partial f}{\partial x}\right)\Bigg|_{(x_0,y_0)}(x - x_0) + \left(\frac{\partial f}{\partial y}\right)\Bigg|_{(x_0,y_0)}(y - y_0).$$

Geometric interpretation of the total differential. The substitution $x - x_0 = \Delta x$ and $y - y_0 = \Delta y$ reduces the above equation to

$$z - z_0 = \left(\frac{\partial f}{\partial x}\right)\Bigg|_{(x_0,y_0)}\Delta x + \left(\frac{\partial f}{\partial y}\right)\Bigg|_{(x_0,y_0)}\Delta y.$$

The expression on the right is the total differential of the function $z = f(x, y)$ at the point $M_0(x_0, y_0)$ in the xy-plane so that $z - z_0 = dz$.

Therefore the total differential of the function $z = f(x, y)$ of two independent variables x and y at the point M_0 corresponding to the increments Δx and Δy is equal to the increment $z - z_0$ of the applicate z of the tangent plane to the surface at the point $P_0(x_0, y_0, f(x_0, y_0))$ obtained in moving from $M_0(x_0, y_0)$ to $M(x_0 + \Delta x, y_0 + \Delta y)$.

Definition. The line that passes through the point $P_0(x_0, y_0, z_0)$ of the surface $F(x, y, z) = 0$ perpendicularly to the tangent plane to this surface at P_0 is called the *normal* to the surface at P_0.

It is easy to notice that the vector $\mathbf{n} = \left(\dfrac{\partial F}{\partial x}, \dfrac{\partial F}{\partial y}, \dfrac{\partial F}{\partial z}\right\}\Bigg|_P$ is the *direction vector* of the normal given by the equations

$$\frac{x - x_0}{\left(\dfrac{\partial F}{\partial x}\right)\Bigg|_{(x_0,\, y_0,\, z_0)}} = \frac{y - y_0}{\left(\dfrac{\partial F}{\partial y}\right)\Bigg|_{(x_0,\, y_0,\, z_0)}} = \frac{z - z_0}{\left(\dfrac{\partial F}{\partial z}\right)\Bigg|_{(x_0,\, y_0,\, z_0)}}.$$

If the surface S is specified by the equation $z = f(x, y)$ the equations of the normal at the point $P_0(x_0, y_0, z_0)$ become

$$\frac{x - x_0}{\left(\frac{\partial f}{\partial x}\right)\Big|_{(x_0, y_0)}} = \frac{y - y_0}{\left(\frac{\partial f}{\partial y}\right)\Big|_{(x_0, y_0)}} = \frac{z - z_0}{-1}.$$

Example. Write down the equations of the tangent plane and the normal to the surface $z = x^2 + y^2$ at the point $O(0, 0, 0)$.

◄ We have $f(x, y) = x^2 + y^2$ so that $\frac{\partial f}{\partial x} = 2x$ and $\frac{\partial f}{\partial y} = 2y$; these derivatives vanish at the point (0, 0), i.e., $f'_x(0, 0) = f'_y(0, 0) = 0$. Then the equation of the tangent plane becomes

$$z - 0 = 0(x - 0) + 0(y - 0),$$

i.e., $z = 0$ (the tangent plane is the xy-plane).

The equations of the normal are

$$\frac{x - 0}{0} = \frac{y - 0}{0} = \frac{z - 0}{1} \quad \text{or} \quad \begin{cases} x = 0 \\ y = 0 \end{cases}$$

i.e., the normal is the z-axis. ►

12.7 Derivatives and Differentials of Higher Orders

Let a function $z = f(x, y)$ have partial derivatives $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$ at every point in some domain D. Clearly, these derivatives $\frac{\partial z}{\partial x} = f'_x(x, y)$ and $\frac{\partial z}{\partial y} = f'_y(x, y)$ are functions of x and y in D; they can have derivatives at some or all points of D.

The partial derivatives of $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$, if they exist, are called the *second derivatives* of the function $z = f(x, y)$ or the *partial derivatives of the second order.* Given a function $z = f(x, y)$ of two independent variables x and y we can write partial derivatives of the second order as

$$\frac{\partial}{\partial x}\left(\frac{\partial z}{\partial x}\right) = \frac{\partial^2 z}{\partial x^2} = f''_{xx}, \quad \frac{\partial}{\partial y}\left(\frac{\partial z}{\partial x}\right) = \frac{\partial^2 z}{\partial y\, \partial x} = f''_{xy},$$

$$\frac{\partial}{\partial x}\left(\frac{\partial z}{\partial y}\right) = \frac{\partial^2 z}{\partial x\, \partial y} = f''_{yx}, \quad \frac{\partial}{\partial y}\left(\frac{\partial z}{\partial y}\right) = \frac{\partial^2 z}{\partial y^2} = f''_{yy}.$$

The derivatives f''_{xy} and f''_{yx} are called the *mixed partial derivatives*; the former is computed by differentiating the given function first with respect to x and then with respect to y and the latter is computed by differentiating the function first with respect to y and then to x.

The partial derivatives of the third and higher orders can be defined in a similar way.

Example. Compute the first and second derivatives of the function $z = x^3y^2 - xy^3$.

◄ We have

$$\frac{\partial z}{\partial x} = 3x^2y^2 - y^3, \ \frac{\partial z}{\partial y} = 2x^3y - 3xy^2,$$

$$\frac{\partial^2 z}{\partial x^2} = 6xy^2, \ \frac{\partial^2 z}{\partial y^2} = 2x^3 - 6xy,$$

$$\frac{\partial^2 z}{\partial y\,\partial x} = 6x^2y - 3y^2, \ \frac{\partial^2 z}{\partial x\,\partial y} = 6x^2y - 3y^2. \ ►$$

Notice that the mixed partial derivatives z''_{xy} and z''_{yx} are identically equal. The explanation of this result is given by the following theorem.

Theorem 12.10. *Let a function $z = f(x, y)$ have partial derivatives f'_x, f'_y, f''_{xy} and f''_{yx} in some neighbourhood of a point $M_0(x_0, y_0)$ and let f''_{xy} and f''_{yx} be continuous at $M_0(x_0, y_0)$.*

Then $f''_{xy}(x_0, y_0) = f''_{yx}(x_0, y_0)$ at M_0.

It is important that the mixed partial derivatives f''_{xy} and f''_{yx} be continuous at $M_0(x_0, y_0)$. For example, the function

$$f(x, y) = \begin{cases} xy\dfrac{x^2 - y^2}{x^2 + y^2}, & x^2 + y^2 \neq 0, \\ 0, & x = y = 0 \end{cases}$$

have mixed partial derivatives f''_{xy} and f''_{yx} which are not continuous at the point $O(0, 0)$ so that $f''_{xy}(0, 0) = -1$ and $f''_{yx}(0, 0) = 1$.

In general, the mth mixed partial derivatives ($m \geqslant 2$) of a function $u = f(x_1, x_2, \ldots, x_n)$ have the same value at a given point if they are continuous at this point and are distinct from each other by the order in which the function is differentiated with respect to its variables.

Differentials of higher orders. Let $z = f(x, y)$ be a function of variables x and y defined on the domain D.

If $z = f(x, y)$ is differentiable on the domain D then the total differential of $z = f(x, y)$ at the point $(x, y) \in D$ corresponding to the increments dx and dy in the variables x and y is given by the formula

$$dz = \frac{\partial z}{\partial x}\,dx + \frac{\partial z}{\partial y}\,dy,$$

where $dx = \Delta x$ and $dy = \Delta y$ are arbitrary increments in independent variables, i.e., these are arbitrary numbers independent of x and y. So we can vary x and y keeping dx and dy constant. Given dx and dy the total differential dz is a function of x and y, which can also be differentiable.

Definition. The total differential of dz at a point (x, y) corresponding to the increments of independent variables equal to dx and dy is called the *second differential* or the *differential of the second order* of a function $z = f(x, y)$ and is denoted by the symbol d^2z so that $d^2z = d(dz)$.

Let a function $z = f(x, y)$ have continuous partial derivatives of the first and second orders in the domain D. Then the total differential dz of $z = f(x, y)$ is a differentiable function so that in D there exists d^2z.

Recall that dx and dy are constant. Then using the laws of differentiation, we obtain

$$d^2z=d(dz)=d\left(\frac{\partial z}{\partial x}\,dx+\frac{\partial z}{\partial y}\,dy\right)=d\left(\frac{\partial z}{\partial x}\,dx\right)+d\left(\frac{\partial z}{\partial y}\,dy\right)$$

$$=d\left(\frac{\partial z}{\partial x}\right)dx+d\left(\frac{\partial z}{\partial y}\right)dy.$$

Applying the formula for the total differential to $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$, we get

$$d\left(\frac{\partial z}{\partial x}\right)=\frac{\partial}{\partial x}\left(\frac{\partial z}{\partial x}\right)dx+\frac{\partial}{\partial y}\left(\frac{\partial z}{\partial x}\right)dy=\frac{\partial^2 z}{\partial x^2}\,dx+\frac{\partial^2 z}{\partial y\,\partial x}\,dy,$$

$$d\left(\frac{\partial z}{\partial y}\right)=\frac{\partial}{\partial x}\left(\frac{\partial z}{\partial y}\right)dx+\frac{\partial}{\partial y}\left(\frac{\partial z}{\partial y}\right)dy=\frac{\partial^2 z}{\partial x\,\partial y}\,dx+\frac{\partial^2 z}{\partial y^2}\,dy,$$

Substituting these relations into the formula for d^2z, we arrive at

$$d^2z=\frac{\partial^2 z}{\partial x^2}(dx)^2+\frac{\partial^2 z}{\partial y\,\partial x}\,dx\,dy+\frac{\partial^2 z}{\partial x\,\partial y}\,dx\,dy+\frac{\partial^2 z}{\partial y^2}(dy)^2.$$

Since $\frac{\partial^2 z}{\partial y\,\partial x}=\frac{\partial^2 z}{\partial x\,\partial y}$ and both mixed partial derivatives are continuous functions we finally obtain

$$d^2z=\frac{\partial^2 z}{\partial x^2}\,dx^2+2\,\frac{\partial^2 z}{\partial x\,\partial y}\,dx\,dy+\frac{\partial^2 z}{\partial y^2}\,dy^2,$$

where $dx^2 = (dx)^2$ and $dy^2 = (dy)^2$.

It is convenient to write the above formula by applying the notation $\frac{\partial}{\partial x}\,dx+\frac{\partial}{\partial y}\,dy$; then we can write

$$d^2z=\left(\frac{\partial}{\partial x}\,dx+\frac{\partial}{\partial y}\,dy\right)^2 z.$$

The symbols $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ are regarded as some "multipliers" so that by completing the square as usually done in algebra, we arrive at the desired result. By way of illustration we write

$$d^2 = \left(\frac{\partial}{\partial x} dx + \frac{\partial}{\partial y} dy\right)^2 = \frac{\partial^2}{\partial x^2} dx^2 + 2\frac{\partial^2}{\partial x \partial y} dx\, dy + \frac{\partial^2}{\partial y^2} dy^2.$$

Multiplying both sides termwise by z and inserting z into the nominators of the "fractions" on the right, we get the formula previously deduced, namely

$$d^2 z = \frac{\partial^2 z}{\partial x^2} dx^2 + 2\frac{\partial^2 z}{\partial x \partial y} dx\, dy + \frac{\partial^2 z}{\partial y^2} dy^2.$$

Formulas for differentials of the third, fourth, etc., orders can be derived in a similar way. In general, the total differential of the nth order denoted by $d^n z$ is the total differential of the total $(n-1)$th differential so that $d^n z = d(d^{n-1} z)$.

If $z = f(x, y) \in C^n(D)$ has continuous partial derivatives up to the nth order derivatives then there exists the total differential of the nth order given by the formula

$$d^n z = \left(\frac{\partial}{\partial x} dx + \frac{\partial}{\partial y} dy\right)^n z.$$

The total nth differential of a function $u = f(x_1, x_2, \ldots, x_m)$ of m independent variables $x_1, x_2, \ldots, x_m$ is given by the formula

$$d^n u = \left(\frac{\partial}{\partial x_1} dx_1 + \frac{\partial}{\partial x_2} dx_2 + \ldots + \frac{\partial}{\partial x_m} dx_m\right)^n u$$

provided that the appropriate conditions are satisfied.

Remark. If x and y are not independent variables but some functions ξ and η then the form of the second differential does not remain invariant. ◄ Indeed, let $z = f(x, y)$ where $x = \varphi(\xi, \eta)$ and $y = \psi(\xi, \eta)$; then the first differential can be written as

$$dz = \frac{\partial z}{\partial x} dx + \frac{\partial z}{\partial y} dy,$$

where dx and dy are some functions and, hence, they cannot be constant. So

$$d^2 z = d\left(\frac{\partial z}{\partial x}\right) dx + d\left(\frac{\partial z}{\partial y}\right) dy + \frac{\partial z}{\partial x} d(dx) + \frac{\partial z}{\partial y} d(dy)$$

$$= \left(\frac{\partial}{\partial x} dx + \frac{\partial}{\partial y} dy\right)^2 z + \frac{\partial z}{\partial x} d^2 x + \frac{\partial z}{\partial y} d^2 y.$$

Thus, the form of the second differential is not invariant in this case. ►

12.8 Taylor's Theorem

Let a function $z = f(x, y)$ have continuous partial derivatives up to the nth derivatives at every point (x, y) in some δ-neighbourhood of the point (x_0, y_0) and let a point $(x_0 + \Delta x, y_0 + \Delta y)$ be contained in the δ-neighbourhood of (x_0, y_0) (Fig. 12.15). We put

$$x = x_0 + t\,\Delta x,$$

$$y = y_0 + t\,\Delta y,$$

where t is some independent variable. Then

$$z = f(x, y) = f(x_0 + t\,\Delta x,\ y_0 + t\,\Delta y) = \varphi(t)$$

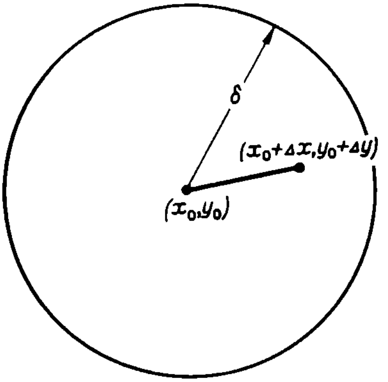


Fig. 12.15

becomes a composite function of t defined on the closed interval $[0, 1]$ and having derivatives up to the nth order on $[0, 1]$. This means that the function $z = \varphi(t)$ can be represented by Taylor's formula in powers of t so that

$$\varphi(t) = \varphi(0) + \frac{\varphi'(0)}{1!}t + \frac{\varphi''(0)}{2!}t^2 + \dots + \frac{\varphi^{(n-1)}(0)}{(n-1)!}t^{n-1} + \frac{\varphi^{(n)}(\theta t)}{n!}t^n,$$

where $0 < \theta < 1$.

Put $t = 1$. Then

$$\varphi(1) = \varphi(0) + \frac{\varphi'(0)}{1!} + \frac{\varphi''(0)}{2!} + \dots + \frac{\varphi^{(n-1)}(0)}{(n-1)!} + \frac{\varphi^{(n)}(\theta)}{n!}, \quad (*)$$

where $0 < \theta < 1$.

We shall express the right-hand side of the above formula in terms of the functions $f(x, y)$ and its derivatives. Notice that being the functions of t the variables x and y of $f(x, y)$ have constant differentials $dx = \Delta x\, dt$ and $dy = \Delta y\, dt$ where Δx and Δy are some given numbers. This means that we can compute differentials of the functions $z = f(x, y)$ by using the formula

$$d^p z = \left(\frac{\partial}{\partial x}\, dx + \frac{\partial}{\partial y}\, dy\right)^p f(x, y) = \left(\frac{\partial}{\partial x}\,(\Delta x\, dt) + \frac{\partial}{\partial y}\,(\Delta y\, dt)\right)^p$$

$$\times f(x, y) = \left(\frac{\partial}{\partial x}\,\Delta x + \frac{\partial}{\partial y}\,\Delta y\right)^p f(x, y)\,(dt)^p,$$

so that

$$\frac{d^p z}{dt^p} = \varphi^{(p)}(t) = \left(\frac{\partial}{\partial x}\,\Delta x + \frac{\partial}{\partial y}\,\Delta y\right)^p f(x, y).$$

Put $t = 0$. Then $x = x_0 + t\,\Delta x = x_0$ and $y = y_0 + t\,\Delta y = y_0$. The above formula yields

$$\varphi^{(p)}_{(0)} = \left(\frac{\partial}{\partial x}\,\Delta x + \frac{\partial}{\partial y}\,\Delta y\right)^p f(x, y)\,\Bigg|_{\substack{x = x_0,\\ y = y_0}} \quad p = 0,\ 1,\ \ldots,\ n - 1.$$

We put $t = \theta$. Then

$$\varphi^{(n)}(\theta) = \left(\frac{\partial}{\partial x}\,\Delta x + \frac{\partial}{\partial y}\,\Delta y\right)^n f(x, y)\,\Bigg|_{\substack{x = x_0 + \theta\,\Delta x\cdot\\ y = y_0 + \theta\,\Delta y}}$$

Also notice that $\varphi(1) = f(x_0 + \Delta x,\ y_0 + \Delta y)$.

Substituting the relations for $\varphi^{(p)}(0)$, $\varphi^{(n)}(\theta)$ and $\varphi(1)$ into (*), we get

$$f(x_0 + \Delta x,\ y_0 + \Delta y) = f(x_0,\ y_0) + \left(\frac{\partial}{\partial x}\,\Delta x + \frac{\partial}{\partial y}\,\Delta y\right) f(x, y)\,\Bigg|_{\substack{x = x_0\\ y = y_0}}$$

$$+ \frac{1}{2!}\left(\frac{\partial}{\partial x}\,\Delta x + \frac{\partial}{\partial y}\,\Delta y\right)^2 f(x, y)\,\Bigg|_{\substack{x = x_0\\ y = y_0}}$$

$$+ \ldots + \frac{1}{(n - 1)!}\left(\frac{\partial}{\partial x}\,\Delta x + \frac{\partial}{\partial y}\,\Delta y\right)^{n-1} f(x, y)\,\Bigg|_{\substack{x = x_0\\ y = y_0}}$$

$$+ \frac{1}{n!}\left(\frac{\partial}{\partial x}\,\Delta x + \frac{\partial}{\partial y}\,\Delta y\right)^n f(x, y)\,\Bigg|_{\substack{x = x_0 + \theta\Delta x,\\ y = y_0 + \theta\Delta y}}$$

where $0 < \theta < 1$.

This is *Taylor's formula for a function* $z = f(x, y)$ *of two variables* and

$$R_n = \frac{1}{n!}\left(\frac{\partial}{\partial x}\Delta x + \frac{\partial}{\partial y}\Delta y\right)^n f(x, y)\Bigg|_{\substack{x = x_0 + \theta\Delta x \\ y = y_0 + \theta\Delta y}}$$

is the *remainder as given by Lagrange.*

To abbreviate the notation it is often convenient to denote $f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0)$ by $\Delta f\,|_{(x_0,y_0)}$. Then Taylor's formula becomes

$$\Delta f\,|_{(x_0,y_0)} = df\,|_{(x_0,y_0)} + \frac{1}{2!}d^2f\,|_{(x_0,y_0)} + \ldots + \frac{1}{(n-1)!}d^{n-1}f\,|_{(x_0,y_0)} + \frac{1}{n!}d^nf\,\Big|_{(x_0 + \theta\Delta x,\ y_0 + \theta\Delta y)}$$

This formula is frequently used as the approximation to the increment Δf of the function $z = f(x, y)$ at the point $M_0(x_0, y_0)$.

If the absolute values of Δx and Δy are sufficiently small and $df \neq 0$ the differential df can be regarded as the approximation to Δf so that $\Delta f \simeq df$. In this case the above formula involves only one term. When this approximation seems to be inadequate the desired precision is attainable by computing some other terms on the right.

Example. Expand the function $f(x, y) = e^x \sin y$ by applying Maclaurin's formula with the remainder of order three.

◀ Taylor's formula with the remainder R_3 takes the form

$$\begin{aligned} f(x_0 + \Delta x, y_0 + \Delta y) &= f(x_0, y_0) + f'_x(x_0, y_0)\Delta x + f'_y(x_0, y_0)\Delta y \\ &+ \frac{1}{2!}\left[f''_{xx}(x_0, y_0)\Delta x^2 + 2f''_{xy}(x_0, y_0)\Delta x\Delta y + f''_{yy}(x_0, y_0)\Delta y^2\right] \\ &+ \frac{1}{3!}\left[f'''_{xxx}(x, y)\Delta x^3 + 3f'''_{xxy}(x, y)\Delta x^2\Delta y + 3f'''_{xyy}(x, y)\Delta x\Delta y^2\right. \\ &\left.+ f'''_{yyy}(x, y)\Delta y^3\right]\Bigg|_{\substack{x = x_0 + \theta\Delta x \\ y = y_0 + \theta\Delta y}} \cdot \end{aligned}$$

Putting $x_0 = y_0 = 0$, $\Delta x = x$ and $\Delta y = y$, we get Maclaurin's formula

$$\begin{aligned} f(x, y) &= f(0, 0) + f'_x(0, 0)x + f'_y(0, 0)y \\ &+ \frac{1}{2!}\left[f''_{xx}(0, 0)x^2 + 2f''_{xy}(0, 0)xy + f''_{yy}(0, 0)y^2\right] \\ &+ \frac{1}{3!}\left[f'''_{xxx}(\theta x, \theta y)x^3 + 3f'''_{xxy}(\theta x, \theta y)x^2y\right. \\ &\left.+ f'''_{xyy}(\theta x, \theta y)xy^2 + f'''_{yyy}(\theta x, \theta y)y^3\right], \quad 0 < \theta < 1. \end{aligned}$$

Recall that $f(x, y) = e^x \sin y$; then

$$
\begin{aligned}
&f(x, y) = e^x \sin y, && f(0, 0) = 0;\\
&f'_x(x, y) = e^x \sin y, && f'_x(0, 0) = 0;\\
&f'_y(x, y) = e^x \cos y, && f'_y(0, 0) = 1;\\
&f''_{xx}(x, y) = e^x \sin y, && f''_{xx}(0, 0) = 0;\\
&f''_{xy}(x, y) = e^x \cos y, && f''_{xy}(0, 0) = 1;\\
&f''_{yy}(x, y) = -e^x \sin y, && f''_{yy}(0, 0) = 0;\\
&f'''_{xxx}(x, y) = e^x \sin y, && f'''_{xxx}(\theta x, \theta y) = e^{\theta x} \sin \theta y;\\
&f'''_{xxy}(x, y) = e^x \cos y, && f'''_{xxy}(\theta x, \theta y) = e^{\theta x} \cos \theta y;\\
&f'''_{xyy}(x, y) = -e^x \sin y, && f'''_{xyy}(\theta x, \theta y) = -e^{\theta x} \sin \theta y;\\
&f'''_{yyy}(x, y) = -e^x \cos y, && f'''_{yyy}(\theta x, \theta y) = -e^{\theta x} \cos \theta y;
\end{aligned}
$$

and Maclaurin's formula gives

$$
\begin{aligned}
e^x \sin y = y + xy + \frac{1}{6}\,[&e^{\theta x} \sin \theta y \cdot x^3 + 3e^{\theta x} \cos \theta y \cdot x^2 y\\
&- 3e^{\theta x} \sin \theta y \cdot xy^2 - e^{\theta x} \cos \theta y \cdot y^3]. \quad ►
\end{aligned}
$$

Remark. It is easy to notice that Maclaurin's formula admits a representation of the form

$$f(x, y)=f(0, 0) + P_1(x, y) + P_2(x, y) + \ldots + P_{n-1}(x, y) + R_n,$$

where $P_k(x, y)$ is a homogeneous polynomial of degree k in x and y.

Sometimes Maclaurin's formula for a given function $f(x, y)$ can easily be obtained by applying asymptotic relations for infinitesimals.

Example. Expand the function $f(x, y) = \dfrac{1}{(1 - x)(1 - y)}$ by applying Maclaurin's formula with the remainder as given by Peano.

◄ We have

$$
\begin{aligned}
f(x, y) &= \frac{1}{(1 - x)(1 - y)} = \frac{1}{1 - x} \cdot \frac{1}{1 - y}\\
&= (1 + x + x^2 + o(x^2))(1 + y + y^2 + o(y^2))\\
&= 1 + x + y + x^2 + xy + y^2 + o(\varrho^2),
\end{aligned}
$$

where $\varrho^2 = x^2 + y^2$. ►

12.9 Extrema of a Function of Several Variables

Let a function $z = f(x, y)$ be defined on some region D and a point $M_0(x_0, y_0)$ be an internal point in this region.

Definition. If there exists a number $\delta > 0$ such that for all Δx and Δy meeting the conditions $|\Delta x| < \delta$ and $|\Delta y| < \delta$ we have

$$\Delta f(x_0, y_0) = f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0) \leqslant 0,$$

then $M_0(x_0, y_0)$ is called the point of a *local maximum* $f(x, y)$; if for all Δx, Δy meeting the conditions $|\Delta x| < \delta$ and $|\Delta y| < \delta$

$$\Delta f(x_0, y_0) = f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0) \geqslant 0,$$

then $M_0(x_0, y_0)$ is called the point of a *local minimum*.

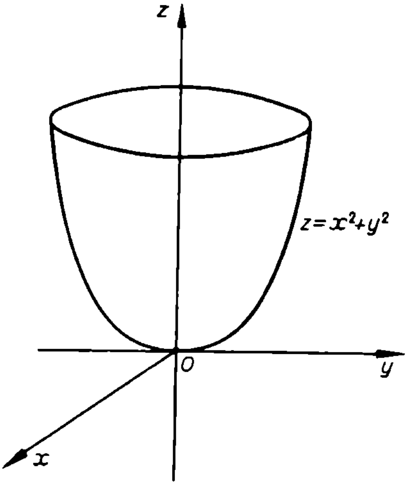


Fig. 12.16

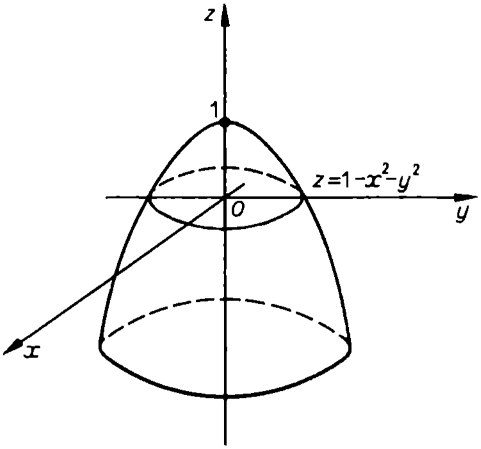


Fig. 12.17

In other words, the point $M_0(x_0, y_0)$ is the point of a maximum or a minimum of $f(x, y)$, if there exists a δ-neighbourhood of $M_0(x_0, y_0)$ such that at all points $M(x, y)$ within this neighbourhood the increment

$$\Delta f = f(x, y) - f(x_0, y_0)$$

retains its sign.

Examples. (1) For the function $z = x^2 + y^2$ the point $O(0, 0)$ is a point of minimum (Fig. 12.16).

(2) For the function $z = 1 - x^2 - y^2$ the point $O(0, 0)$ is the point of maximum (Fig. 12.17).

(3) For the function

$$f(x, y) = \begin{cases} x^2 + y^2, & x^2 + y^2 \neq 0, \\ 1, & x = y = 0, \end{cases}$$

the point $O(0, 0)$ is the point of maximum (Fig. 12.18).

◀ There exists a neighbourhood of the point $O(0, 0)$, e.g., a circle of radius 1/2 (see Fig. 12.18), such that at all points in it other than $O(0, 0)$ the value of the function $f(x, y)$ is less than $1 = f(0, 0)$. ▶

We will only consider points of strict maximum and minimum, i.e., ones for which the strict inequalities $\Delta f < 0$ or $\Delta f > 0$ are obeyed, respectively, for all the points $M(x, y)$ in some δ-neighbourhood of the point M_0 other than M_0.

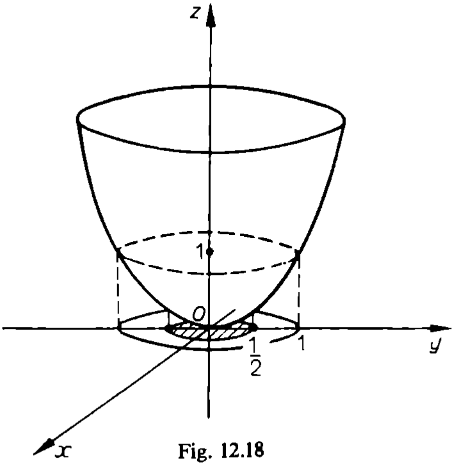


Fig. 12.18

The value of a function at a maximum point is called a *maximum*, and the value of the function at a minimum point is called a *minimum* of the function. Maximum and minimum points of the function are called *extremum* points of the function, and the maxima and minima themselves are called its *extrema*.

Theorem 12.11 (necessary condition for extremum). *If a function $z = f(x, y)$ has an extremum at a point $M_0(x_0, y_0)$, then at that point each partial derivative $\partial z/\partial x$ and $\partial z/\partial y$ either vanishes or does not exist.*

◀ Suppose that the function $z = f(x, y)$ has an extremum at the point $M_0(x_0, y_0)$. We will assign to the variable y a value y_0. The function $z = f(x, y)$ will then be a function of one variable x: $z = f(x, y_0)$. Since at $x = x_0$ it has an extremum (maximum or minimum; Fig. 12.19), its derivative with respect to z at $x = x_0$, i.e., $(\partial z/\partial x)\,|_{(x_0,y_0)}$ either vanishes or does not exist. Likewise, we see that $\partial z/\partial y\,|_{(x_0,y_0)}$ is zero or does not exist. ▶

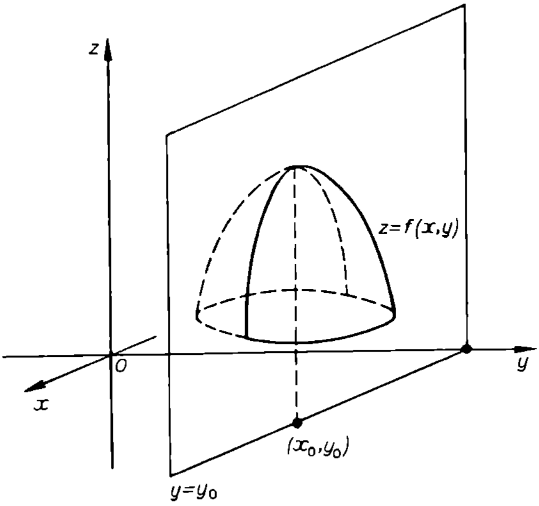


Fig. 12.19

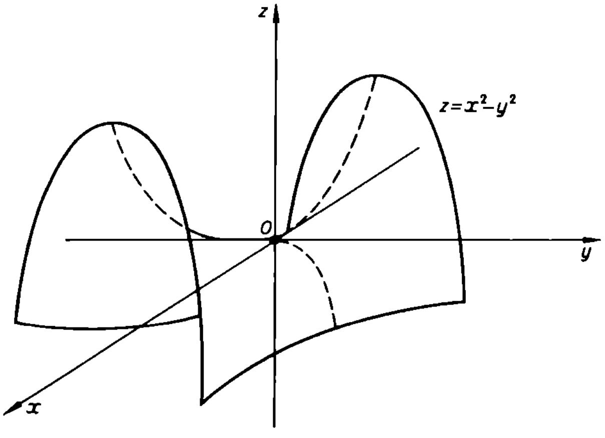


Fig. 12.20

Points where $\partial z/\partial x$ and $\partial z/\partial y$ are zero or do not exist are called the *critical points* of the function $z = f(x, y)$, points where $\partial z/\partial x$ and $\partial z/\partial y$ are zero are called its stationary points.

The theorem only gives the necessary conditions for an extremum. For example the function $z = x^2 - y^2$ has derivatives $\partial z/\partial x = 2x$, $\partial z/\partial y = -2y$, which become zero at $x = y = 0$. But at $O(0, 0)$ this function has no extremum.

Indeed, the function $f(x, y) = x^2 - y^2$ vanishes at the point $O(0, 0)$ and at the point $M(x, y)$ that are as close as possible to $O(0, 0)$, it assumes positive or negative values. For it

$$\Delta f(0, 0) = f(x, y) - f(0, 0) = x^2 - y^2 \text{ and } \begin{cases} \Delta f > 0 \text{ at points } (x, 0) \\ \Delta f < 0 \text{ at points } (0, y) \end{cases}$$

at arbitrary small $|x| > 0$ and $|y| > 0$.

Point $O(0, 0)$ of this type is called a *minimax point* (Fig. 12.20).

Sufficient conditions for an extremum of a function of two variables are given by the following theorem:

Theorem 12.12. *Let a point* $M_0(x_0, y_0)$ *be a stationary point of a function* $f(x, y)$, *i.e.,*

$$f'_x(x_0, y_0) = 0 \text{ and } f'_y(x_0, y_0) = 0.$$

Suppose that at some neighbourhood of the point $M_0(x_0, y_0)$ *including* M_0 *itself, the function* $f(x, y)$ *has continuous partial derivatives up to the second order. Then:*

(i) *the function* $f(x, y)$ *has a maximum at* $M_0(x_0, y_0)$ *if at that point the determinant*

$$D = \begin{vmatrix} f'_{xx}(x_0, y_0) & f''_{xy}(x_0, y_0) \\ f''_{xy}(x_0, y_0) & f''_{yy}(x_0, y_0) \end{vmatrix}$$

$$= f''_{xx}(x_0, y_0) f''_{yy}(x_0, y_0) - f''^{\,2}_{xy}(x_0, y_0) > 0$$

and $f''_{xx}(x_0, y_0) < 0$ $(f''_{yy}(x_0, y_0) < 0)$;

(ii) *the function* $f(x, y)$ *has a minimum at* $M_0(x_0, y_0)$ *if*

$$D = f''_{xx}(x_0, y_0) f''_{yy}(x_0, y_0) - f''^{\,2}_{xy}(x_0, y_0) > 0$$

and $f''_{xx}(x_0, y_0) > 0$ $(f''_{yy}(x_0, y_0) > 0)$;

(iii) *the function* $f(x, y)$ *has no extremum at* $M_0(x_0, y_0)$ *if*

$$D = f''_{xx}(x_0, y_0) f''_{yy}(x_0, y_0) - f''^{\,2}_{xy}(x_0, y_0) > 0.$$

If $D = 0$, *then at the point* $M_0(x_0, y_0)$ *the function* $f(x, y)$ *may or may not have an extremum.* Further examination is then needed.

◄ We will only prove items (i) and (ii) of the theorem. We write Taylor's formula of the second order for $f(x, y)$:

$$f(x_0 + \Delta x, y_0 + \Delta y) = f(x_0, y_0) + f'_x(x_0, y_0)\,\Delta x$$
$$+ f'_y(x_0, y_0)\Delta y + \frac{1}{2}\,[f''_{xx}(x, y)\Delta x^2 + 2f''_{xy}(x, y)\,\Delta x\,\Delta y$$
$$+ f''_{yy}(x, y)\Delta y^2]\,\bigg|_{\substack{x = x_0 + \theta\Delta x, \\ y = y_0 + \theta\Delta y}}$$

where $0 < \theta < 1$.

As stated $f'_x(x_0, y_0) = 0$, $f'_y(x_0, y_0) = 0$, so that

$$\Delta f = f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0)$$
$$= \frac{1}{2}\left[f''_{xx}(x, y)\Delta x^2 + 2f''_{yy}(x, y)\Delta x\,\Delta y + f''_{yy}(x, y)\Delta y^2\right]\Bigg|_{\substack{x = x_0 + \theta\Delta x \\ y = y_0 + \theta\Delta y}} \cdot \quad (*)$$

It is seen that the sign of the increment Δf is determined by that of the trinomial on the right of $(*)$, i.e., by the sign of the second differential d^2f.

We denote for short

$$A = f''_{xx}(x, y),\ B = f''_{xy}(x, y),\ C = f''_{yy}(x, y)$$

and write $(*)$ in the form

$$\Delta f = \frac{1}{2}(A\,\Delta x^2 + 2B\,\Delta x\,\Delta y + C\,\Delta y^2)\Bigg|_{\substack{x = x_0 + \theta\Delta x \\ y = y_0 + \theta\Delta y}} \cdot \qquad (**)$$

Suppose that at $M_0(x_0, y_0)$ we have $AC - B^2 > 0$, i.e.,

$$f''_{xx}(x_0, y_0) - f''_{yy}(x_0, y_0) - \ f''^{\,2}_{xy}(x_0, y_0) > 0.$$

Since the partial derivatives of the second order of $f(x, y)$ are continuous, the inequality $AC - B^2 > 0$ will hold within some neighbourhood of $M_0(x_0, y_0)$.

When $AC - B^2 > 0$, $A = f''_{xx}(x, y) \neq 0$ at the point $M_0(x_0, y_0)$, and therefore within some neighbourhood of M_0, since $f''_{xx}(x, y)$ is continuous, it has a constant sign that coincides with that of A at (x_0, y_0).

But in the region where $A \neq 0$, we have

$$A\,\Delta x^2 + 2B\,\Delta x\,\Delta y + C\,\Delta y^2 = \frac{1}{A}\left[(A\,\Delta x + B\,\Delta y)^2 + (AC - B^2)\Delta y^2\right].$$

It is seen that for $AC - B^2 > 0$ in some neighbourhood of the point $M_0(x_0, y_0)$ the sign of the trinomial $A\,\Delta x^2 + 2B\,\Delta x\,\Delta y + C\,\Delta y^2$ coincides with the sign of A at the point (x_0, y_0) (and also with the sign of C, because A and C cannot have unlike signs for $AC - B^2 > 0$).

Since by virtue of $(**)$ the sign of $A\,\Delta x^2 + 2B\,\Delta x\,\Delta y + C\,\Delta y^2$ at the point $(x_0 + \theta\,\Delta x, y_0 + \theta\,\Delta y)$ determines the sign of the difference

$$\Delta f = f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0),$$

we come to the following conclusion: if for the function $f(x, y)$ at the stationary point (x_0, y_0) the conditions $AC - B^2 > 0$ and $A < 0$ $(C < 0)$ hold, then for all sufficiently small $|\Delta x|$ and $|\Delta y|$ we have

$$\Delta f = f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0) \leqslant 0,$$

and so at (x_0, y_0) the function has a maximum.

If at the stationary point (x_0, y_0) $AC - B^2 > 0$ and $A > 0$ $(C > 0)$, then for all sufficiently small $|\Delta x|$ and $|\Delta y|$ we have

$$\Delta f = f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0) \geqslant 0,$$

and so at (x_0, y_0) the function $f(x, y)$ has a minimum. ►

Examples. (1) Examine for extremum the function

$$z = x^2 + 2y^2 - 2x + 4y - 6.$$

◄ Using the necessary conditions for an extremum, we find the stationary points. For this purpose, we find the partial derivatives $\partial z/\partial x$ and $\partial z/\partial y$ and equate them to zero.

We obtain the system of equations

$$\begin{cases} \dfrac{\partial z}{\partial x} = 2x - 2 = 0, \\ \dfrac{\partial z}{\partial y} = 4y + 4 = 0. \end{cases}$$

Hence $x = 1$, $y = -1$, so that $M_0(1, -1)$ is a stationary point.

We now make use of the Theorem 12.12.

We have

$$A\Big|_{M_0} = \frac{\partial^2 z}{\partial x^2}\Big|_{M_0} = 2, \quad B\Big|_{M_0} = \frac{\partial^2 z}{\partial x\,\partial y}\Big|_{M_0} = 0,$$

$$C\Big|_{M_0} = \frac{\partial^2 z}{\partial y^2}\Big|_{M_0} = 4,$$

so that $(AC - B^2)\,|_{M_0} = 8 > 0$.

Therefore, at M_0 we have an extremum. Since $A\,|_{M_0} = 2 > 0$, this is a minimum.

If we thansform z to the form

$$z = (x - 1)^2 + 2(y + 1)^2 - 9, \tag{$*$}$$

we can easily see that the right side of $(*)$ will be minimal when $x = 1$, $y = -1$. This is an absolute minimum of the function. ►

(2) Examine for extremum the function $z = xy$.

◄ We find the stationary points of the function, for which purpose we make up the system of equations

$$\begin{cases} \dfrac{\partial z}{\partial x} = y = 0, \\ \dfrac{\partial z}{\partial y} = x = 0. \end{cases}$$

Hence $x = y = 0$, so that $M_0(0, 0)$ is a stationary point. Since

$$A\Big|_{M_0} = \frac{\partial^2 z}{\partial x^2}\Big|_{M_0} = 0, \quad B\Big|_{M_0} = \frac{\partial^2 z}{\partial x\,\partial y}\Big|_{M_0} = 1,$$

$$C\Big|_{M_0} = \frac{\partial^2 z}{\partial y^2}\Big|_{M_0} = 0,$$

$(AC - B^2)|_{M_0} = -1 < 0$, and by Theorem 12.12 at the point $M_0(0, 0)$ there is no extremum. ►

(3) Examine for extremum the function $z = x^4 + y^4$.

◄ We find the stationary points of the function. From the system of equations

$$\begin{cases} \dfrac{\partial z}{\partial x} = 4x^3 = 0, \\ \dfrac{\partial z}{\partial y} = 4y^3 = 0, \end{cases}$$

we get $x = y = 0$, so that the point $M_0(0, 0)$ is a stationary point. Further, we have

$$A\Big|_{M_0} = \frac{\partial^2 z}{\partial x^2}\Big|_{M_0} = 0, \quad B\Big|_{M_0} = \frac{\partial^2 z}{\partial x\,\partial y}\Big|_{M_0} = 0,$$

$$C\Big|_{M_0} = \frac{\partial^2 z}{\partial y^2}\Big|_{M_0} = 0$$

so that $(AC - B^2)\,|_{M_0} = 0$ and the theorem gives us no answer as to whether or not there exists an extremum.

Therefore, we proceed as follows. For the function $z = x^4 + y^4$ at all points $M(x, y)$ other than $M_0(0, 0)$, we will have

$$\Delta f(0, 0) = f(x, y) - f(0, 0) = x^4 + y^4 > 0,$$

so that, by definition, at the point $M_0(0, 0)$ the function z has an absolute minimum.

Reasoning along the same lines, we establish that the function $z = -x^4 - y^4$ at the point $O(0, 0)$ has a maximum, and the function $z = x^4 - y^4$ at $O(0, 0)$ has no extremum. ►

Conditional extremum. Up to this point we have been dealing with local extrema of a function whose arguments are not subject to any additional conditions. Such extrema are known as *unconditional ones.*

However, problems are quite common which deal with conditional extrema. Take a function $z = f(x, y)$ defined on a region D. Suppose that on this region a curve L is defined and we want to find the extrema of $f(x, y)$ only among those values which correspond to the points of the curve L. Such extrema are called *conditional extrema* of the function $z = f(x, y)$ on L.

Definition. We say that at the point $M_0(x_0, y_0)$ lying on L the function $f(x, y)$ has a conditional maximum (minimum), if the inequality $f(x, y) < f(x_0, y_0)$ (or $f(x, y) > f(x_0, y_0)$) holds at all points $M(x, y)$ on the curve L that belong to some neighbourhood of the point $M_0(x_0, y_0)$ and are different from M_0 (Fig. 12.21).

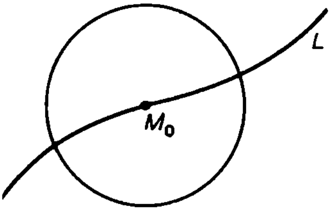


Fig. 12.21

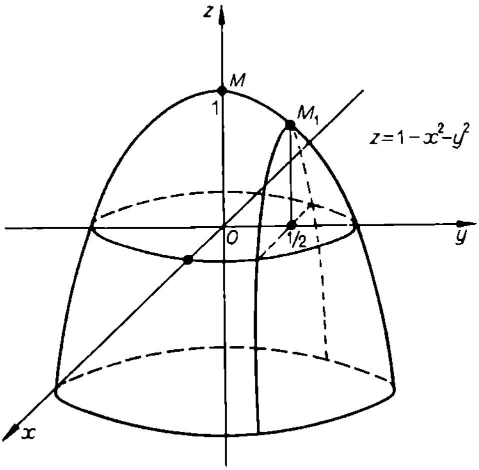


Fig. 12.22

If the equation of the curve L is $\varphi(x, y) = 0$, then the problem on finding a conditional extremum of $z = f(x, y)$ of the curve L can be formulated as follows: find extrema of the function $z = f(x, y)$ in D under the condition that $\varphi(x, y) = 0$.

Therefore, in finding conditional extrema of the function $z = f(x, y)$ the arguments x and y can no longer be regarded as independent variables. They are connected by the relation $\varphi(x, y) = 0$, which is known as a *constraint equation.*

To clarify the difference between unconditional and conditional extrema we will take an example. The unconditional maximum of the function $z = 1 - x^2 - y^2$ (Fig. 12.22) is equal to unity and it is achieved at the point (0, 0). Corresponding to it is the point M, i.e., the vertex of the paraboloid. We now add the constraint equation $y = 1/2$. The conditional maximum will, clearly, be equal to 3/4. It is achieved at the point (0, 1/2), and corresponding to it is the vertex M_1 of the parabola, which is the line of intersection of the paraboloid with the plane $y = 1/2$. In the case of the unconditional maximum we have the maximum z-coordinate among all the applicates of the surface $z = 1 - x^2 - y^2$; in the case of the conditional maximum we have the maximum z-coordinate only among those for the straight line $y = 1/2$ in the xy-plane.

One method of finding a conditional extremum of the function

$$z = f(x, y)$$

with the constraint

$$\varphi(x, y) = 0$$

is as follows.

Suppose now that the constraint equation $\varphi(x, y) = 0$ defines y as a single-valued differentiable function of x: $y = \psi(x)$. Substituting $\psi(x)$ for y in $z = f(x, y)$, we obtain a function of one argument

$$z = f(x, \psi(x)) = F(x),$$

which takes into account the constraint condition.

The extremum (unconditional) of $F(x)$ is the desired conditional extremum.

Example. Find the extremum of the function

$$z = x^2 + y^2 \tag{$*$}$$

subject to

$$x + y - 1 = 0. \tag{$**$}$$

◄ From the constraint equation ($**$) we find $y = 1 - x$. Substituting this value of y into ($*$), we obtain a function of one argument x: $z = x^2 + (1 - x)^2$.

Examine it for extremum: $z' = 2x - 2(1 - x)$, hence $x = 1/2$ is a critical point; $z'' = 4 > 0$, so that $x = 1/2$ ($y = 1/2$) is a conditional minimum of z (Fig. 12.23). ►

There is another method of solving the problem on conditional extremum, called the method of *Lagrange multipliers.*

◄ Let $M_0(x_0, y_0)$ be a point of conditional extremum of a function $z = f(x, y)$ with the constraint $\varphi(x, y) = 0$. If we take y to be $\psi(x)$ defined by the constraint equation $\varphi(x, y) = 0$, we find that the derivative with respect to x of $f(x, y)$ at the point M_0 must be zero or, equivalently, the differential of $f(x, y)$ at M_0 must be zero:

$$(df)\,|_{M_0} = (f'_x dx + f'_y dy)\,|_{M_0} = 0. \tag{$***$}$$

From the constraint equation we have

$$(d\varphi)\,|_{M_0} = (\varphi_x' dx + \varphi_y' dy)\,|_{M_0} = 0.$$

Multiplying the last relation by a numerical factor λ yet to be found and adding up the result term by term with (∗∗∗), we obtain

$$(f_x' + \lambda\varphi_x')\,|_{M_0}\, dx + (f_y' + \lambda\varphi_y')\,|_{M_0}\, dy = 0.$$

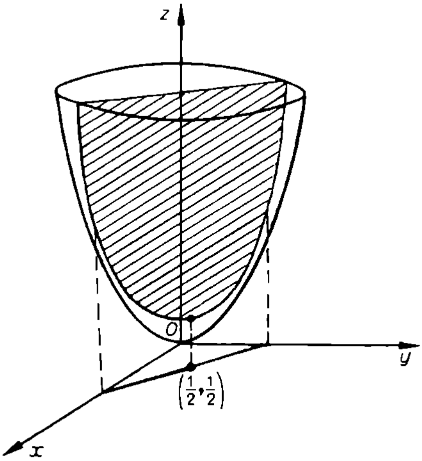


Fig. 12.23

Suppose that λ is taken such that

$$(f_y' + \lambda\varphi_y')\,|_{M_0} = 0$$

(we consider that $\varphi_y' \neq 0$); then, since dx is arbitrary, we have

$$(f_x' + \lambda\varphi_x')\,|_{M_0} = 0.$$

The last two relations reflect the necessary conditions for an unconditional extremum at the point $M_0(x_0, y_0)$ of the function

$$F(x, y) = f(x, y) + \lambda\varphi(x, y),$$

called the *Lagrange function.*

Therefore, a conditional extremum of $f(x, y)$, if $\varphi(x, y) = 0$, is bound to be a stationary point of the Lagrange function

$$F(x, y) = f(x, y) + \lambda\varphi(x, y),$$

where λ is some numerical factor. ►

We thus obtain the following rule to find conditional extrema.

To find points that may be conditional extrema of the function $z = f(x, y)$ subject to the constraint $\varphi(x, y) = 0$, we:

(1) form the Lagrange function

$$F(x, y) = f(x, y) + \lambda \varphi(x, y);$$

(2) equate to zero the derivatives $\partial F/\partial x$ and $\partial F/\partial y$ of the function and, subjecting the resultant equations to the constraint equation, obtain a system of three equations

$$\begin{cases} \dfrac{\partial F}{\partial x} \equiv f_x'(x, y) + \lambda \varphi_x'(x, y) = 0, \\ \dfrac{\partial F}{\partial y} \equiv f_y'(x, y) + \lambda \varphi_y'(x, y) = 0, \\ \dfrac{\partial F}{\partial \lambda} \equiv \varphi(x, y) = 0, \end{cases} \tag{*}$$

from which we find the values λ and the coordinates x and y of the possible extremum points.

The question of the existence and nature of a conditional extremum is solved by examining the sign of the second differential of the Lagrange function

$$d^2F(x, y) = \frac{\partial^2 F}{\partial x^2} dx^2 + 2 \frac{\partial^2 F}{\partial x \partial y} dx\, dy + \frac{\partial^2 F}{\partial y^2} dy^2$$

for the set of values x_0, y_0, λ obtained from (*) subject to the condition $\dfrac{\partial \varphi}{\partial x} dx + \dfrac{\partial \varphi}{\partial y} dy = 0 \ ((dx)^2 + (dy)^2 \neq 0)$.

If $d^2F < 0$, then at the point (x_0, y_0) the function $f(x, y)$ has a conditional maximum, if $d^2F > 0$, a conditional minimum.

Specifically, if at a stationary point (x_0, y_0) the determinant D for $F(x, y)$ is positive:

$$D(x_0, y_0) = \begin{vmatrix} F_{xx}''(x_0, y_0) & F_{xy}''(x_0, y_0) \\ F_{xy}''(x_0, y_0) & F_{yy}''(x_0, y_0) \end{vmatrix} > 0,$$

then at the point (x_0, y_0) $f(x, y)$ has a conditional maximum, if $A = F_{xx}''(x_0, y_0) < 0$, and a conditional minimum if $A = F_{xx}''(x_0, y_0) > 0$.

We now return to the example on p. 574 where we sought for an extremum of the function $z = x^2 + y^2$ subject to the condition that $x + y = 1$. We will solve the problem by the method of the Lagrange multipliers.

◄ The Lagrange function will then be

$$F(x, y; \lambda) = x^2 + y^2 + \lambda(x + y - 1).$$

To find stationary points we will construct the system:

$$\begin{cases} F'_x \equiv 2x + \lambda = 0, \\ F'_y \equiv 2y + \lambda = 0, \\ F'_\lambda \equiv x + y - 1 = 0. \end{cases}$$

From the first two equations of the system we will obtain $x = y$, and then from the third equation of the system (constraint equation) we find $x = y = 1/2$ (the coordinates of a possible extremum). It appears that $\lambda = -1$. The Lagrange function will thus be

$$F(x, y; -1) = x^2 + y^2 - x - y + 1.$$

For it $F''_{xx} = 2$, $F''_{yy} = 2$, $F''_{xy} = 0$, so that $D = \begin{vmatrix} 2 & 0 \\ 0 & 2 \end{vmatrix} = 4 > 0$ and $F''_{xx} = 2 > 0$, i.e., the point $M_0(1/2, 1/2)$ is a conditional minimum of the function $z = x^2 + y^2$ subject to the condition $x + y = 1$. ►

The fact that there is no unconditional extremum for the Lagrange function $F(x, y)$ does not yet imply that there is no conditional extremum for the function $f(x, y)$ in the presence of the constaint $\varphi(x, y) = 0$.

Example. Find an extremum of the function $z = xy$ subject to the condition $y - x = 0$.

◄ We write the Lagrange function

$$F(x, y; \lambda) = xy + \lambda(y - x)$$

and also a system to find λ and the coordinates of possible extrema:

$$\begin{cases} F'_x \equiv y - \lambda = 0, \\ F'_y \equiv x + \lambda = 0, \\ F'_\lambda \equiv y - x = 0. \end{cases} \qquad (**)$$

From the first two equations we find $x + y = 0$ and obtain the system

$$\begin{cases} x + y = 0, \\ y - x = 0, \end{cases}$$

whence $x = y = 0$. Then, $\lambda = 0$. The Lagrange function will thus be

$$F(x, y; 0) = xy.$$

At the point $(0, 0)$ $F(x, y; 0)$ has no unconditional extremum; however, $z = xy$ has a conditional extremum when $y = x$. Indeed, we then have $z = x^2$, and so at the point $(0, 0)$ there is a conditional minimum. ►

The method of the Lagrange multipliers can be extended to include the case of functions of any number of arguments.

We wish to find an extremum of the function

$$z = f(x_1, x_2, \ldots, x_n) \tag{*}$$

subject to the constraint equations

$$\begin{cases} \varphi_1(x_1, x_2, \ldots, x_n) = 0, \\ \varphi_2(x_1, x_2, \ldots, x_n) = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . \\ \varphi_m(x_1, x_2, \ldots, x_n) = 0, \end{cases} \tag{**}$$

where $m < n$.

We form the Lagrange function

$$F(x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n) + \lambda_1\varphi_1(x_1, x_2, \ldots, x_n)$$
$$+ \lambda_2\varphi_2(x_1, x_2, \ldots, x_n) + \ldots + \lambda_m\varphi_m(x_1, x_2, \ldots, x_n),$$

where $\lambda_1, \lambda_2, \ldots, \lambda_m$ are constant multipliers to be found.

Equating to zero all the partial derivatives of the first order of F and attaching to the resultant equations the constraint equations (**), we will obtain a system of $n + m$ equations from which we find $\lambda_1, \lambda_2, \ldots, \lambda_m$ and the coordinates $x_1, x_2, \ldots, x_n$ of the possible points of conditional extremum. The question of whether or not the points obtained by the Lagrange method are conditional extrema can often be solved from physical or geometrical considerations.

Absolute maximum and minimum of a continuous function. We would like to find the largest or smallest value of a function $z = f(x, y)$ (i.e., its absolute maximum or minimum), the function being continuous on some closed region $\overline{D}$. By Theorem 12.3 in this region there is a point (x_0, y_0) at which the function attains an absolute maximum (minimum). If the point (x_0, y_0) lies within the region $\overline{D}$, then the function f has a maximum (minimum) in $\overline{D}$, so that in this case the point of interest is among the critical points of $f(x, y)$. However, the function may also attain its absolute maximum (minimum) at the boundary of the region.

Therefore, to find the absolute maximum (minimum) of the function $z = f(x, y)$ in the bounded closed region $\overline{D}$ we will have to find all maxima (minima) of the function within the region, and also at the boundary of the region. The largest (smallest) of these numbers will be the desired absolute maximum (minimum) of $z = f(x, y)$ in $\overline{D}$. We now turn to a differentiable function.

Example. Find the absolute maximum and minimum of the function $z = x^2 + y^2$ in the region $\overline{D}\ \{-1 \leqslant x \leqslant 1,\ -1 \leqslant y \leqslant 1\}$.

◄ We look for the critical point of $z = x^2 + y^2$ inside $\overline{D}$. To this end, we form up the system of equations

$$\begin{cases} \dfrac{\partial z}{\partial x} \equiv 2x = 0, \\ \dfrac{\partial z}{\partial y} \equiv 2y = 0. \end{cases}$$

Hence, $x = y = 0$, so that the point $O(0, 0)$ is a critical point of z. Since $\dfrac{\partial^2 z}{\partial x^2} = 2$, $\dfrac{\partial^2 z}{\partial y^2} = 2$, $\dfrac{\partial^2 z}{\partial x \partial y} = 0$, at $O(0,\ 0)$ $AC - B^2 = 4 > 0$ and $A = 2 > 0$, and so at this point the function $z = x^2 + y^2$ has a minimum equal to zero.

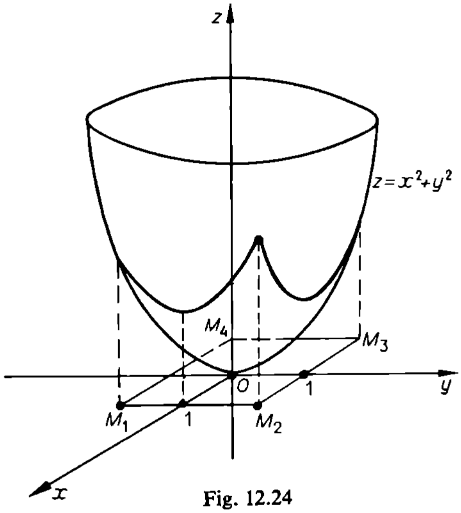


Fig. 12.24

We now find the largest and smallest values of the function on the boundary Γ of $\overline{D}$. On the part of the boundary Γ_1 $\{x = 1,\ -1 \leqslant y \leqslant 1\}$ we have $z = 1 + y^2$, $\dfrac{dz}{dy} = 2y$, so that $y = 0$ is a critical point. Since $\dfrac{d^2 z}{dy^2} = 2 > 0$, at this point the function $z = 1 + y^2$ has a minimum equal to unity.

At the ends of the segment Γ_1, at points $(1, -1)$ and $(1, 1)$ we have $z(1, -1) = z(1, 1) = 2$. Using the considerations of symmetry, we can obtain the same results for other parts of the boundary: Γ_2 $\{y = 1,\ -1 \leqslant x \leqslant 1\}$, Γ_3 $\{x = -1,\ -1 \leqslant y \leqslant 1\}$ and Γ_4 $\{y = 1,\ -1 \leqslant x \leqslant 1\}$.

We finally find that the absolute minimum $z = x^2 + y^2$ in $\overline{D}$ is zero and it is achieved at the internal point $O(0, 0)$ of the region, and the absolute maximum of the function, equal to two, is achieved at four points on the boundary: $M_1(1, -1)$, $M_2(1, 1)$, $M_3(-1, 1)$, $M_4(-1, -1)$ (Fig. 12.24). ►

Exercises

Find the domains of the following functions:

1. $z = \sin^{-1}\frac{x}{2} + \sqrt{xy}$. **2.** $z = \sqrt{1 - x^2} + \sqrt{1 - y^2}$. **3.** $z=\sqrt{\sin(x^2+y^2)}$.

4. $z = \dfrac{1}{x - y} + \dfrac{1}{y}$. **5.** $z = \ln(x^2 + y)$. **6.** $z = xy + \sqrt{x^2 + y^2 - R^2} + \sqrt{\ln\dfrac{R^2}{x^2 + y^2}}$. **7.** $z = \cot\pi(x + y)$. **8.** (a) $z = \sqrt{\sin x \sin y}$; (b) $z = \sqrt{\sin x - 1} + \sqrt{\sin y - 1}$.

Draw the level lines of the functions

9. (a) $z = x + y$; (b) $z = x^2 + y^2$.

10. (a) $z = \dfrac{y}{x^2}$; (b) $z = \dfrac{y}{\sqrt{x}}$.

11. (a) $z = \ln(x^2 + y)$; (b) $z = \sin^{-1}xy$.

Find the level surfaces of the functions of three independent variables:

12. $u = x + y + z$. **13.** $u = x^2 + y^2 - z^2$.

Find the limits of the functions:

14. (a) $\lim\limits_{\substack{x\to 0\\ y\to 0}} \dfrac{2 - \sqrt{xy + 4}}{xy}$; (b) $\lim\limits_{\substack{x\to 0\\ y\to 0}} \dfrac{x^2y}{x^2 + y^2}$.

15. (a) $\lim\limits_{\substack{x\to 0\\ y\to 0}} \dfrac{\sin(xy)}{xy}$; (b) $\lim\limits_{\substack{x\to 0\\ y\to 2}} \dfrac{\sin(xy)}{x}$.

16. (a) $\lim\limits_{\substack{x\to \infty\\ y\to k}} \left(1 + \dfrac{y}{x}\right)^x$; (b) $\lim\limits_{\substack{x\to \infty\\ y\to \infty}} \dfrac{x + y}{x^2 + y^2}$.
(k = const)

17. (a) $\lim\limits_{\substack{x\to 0\\ y\to 0}} \dfrac{x^2 - y^2}{x^2 + y^2}$; (b) $\lim\limits_{\substack{x\to 0\\ y\to 0}} \dfrac{e^{-\frac{1}{x^2+y^2}}}{x^2 + y^2}$.

18. Show that the function $z = \dfrac{x+y}{x-y}$ has no limit as $x \to 0$ and $y \to 0$. Consider the behaviour of the function on the straight lines $y = kx$.

Find the sets of discontinuity points for the following functions.

19. (a) $z = \dfrac{2}{x^2+y^2}$; (b) $z = \ln\sqrt{x^2+y^2}$.

20. (a) $z = \dfrac{1}{1-x^2-y^2}$; (b) $z = \dfrac{1}{(x-y)^2}$.

21. (a) $z = \cos\dfrac{1}{xy}$; (b) $z = xy$.

22. $z = \dfrac{1}{\sin^2 \pi x + \sin^2 \pi y}$.

Find partial derivatives of the functions and their total differentials:

23. $z = x^3 + y^2 - 2xy$.

24. $z = \tan^{-1}\dfrac{x}{y}$.

25. $z = e^{-x/y}$.

26. $z = \ln(x + \ln y)$.

27. $u = xy + yz + xz$.

28. $u = \sqrt{x^2+y^2+z^2}$.

29. $z = \cosh(x^2 y + \sinh y)$.

30. $u = x^{yz}$.

Find the derivatives of the composite functions:

31. (a) $z = x^2 + xy + y^2$, where $x = t^2$, $y = t$. Find $\dfrac{dz}{dt}$.

(b) $z = y/x$, where $x = e^t$, $y = 1 - e^{2t}$. Find $\dfrac{dz}{dt}$.

32. (a) $z = xe^y$, where $y = \tan^{-1} x$. Find $\dfrac{\partial z}{\partial x}$ and $\dfrac{dz}{dx}$.

(b) $z = \ln(x^2 - y^2)$, where $y = e^x$. Find $\dfrac{\partial z}{\partial x}$ and $\dfrac{dz}{dx}$.

33. (a) $z = \tan^{-1}\dfrac{x}{y}$, where $x = u \sin v$, $y = u \cos v$. Find $\dfrac{\partial z}{\partial u}$ and $\dfrac{\partial z}{\partial v}$.

(b) $z = x^2 + y^2$, where $x = u + v$, $y = u - v$. Find $\dfrac{\partial z}{\partial u}$ and $\dfrac{\partial z}{\partial v}$.

Using the formula for the derivative of a composite function of two variables, find $\dfrac{\partial z}{\partial x}$ and $\dfrac{\partial z}{\partial y}$ of the functions:

34. (a) $z = f(u)$, where $u = \sin^{-1} xy + \dfrac{x}{y}$,

(b) $z = f(u)$, where $u = \sin\dfrac{x}{y} + e^{\tan xy}$.

35. (a) $z = f(u, v)$, where $u = x^2 \ln y$, $v = \sin^{-1}\frac{x}{y}$,

(b) $z = f(u, v)$, where $u = e^{x^2 + \cos y}$, $v = \tan^{-1}\frac{y}{x}$.

Find $\frac{dy}{dx}$ of the functions given implicitly by the equations:

36. $x^2 + y^2 + \ln(x^2 + y^2) = a^2$. 37. $\ln\tan\frac{y}{x} - \frac{y}{x} = b$.

38. $x^2y + \sin^{-1}\frac{x}{y} + \frac{1}{y} = 0$. 39. $y^x = x^y$.

40. Find the slope of the tangent to the curve $x^2 + y^2 = 10y$ at its intersection with the straight line $x = 3$.

41. Find the points where the tangent to the curve $x^2 + y^2 + 2x - 2y - 2 = 0$ is parallel to the x-axis.

Find $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$:

42. $x\cos y + y\cos z + z\cos x = 1$.

43. $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$.

Write the equations of the tangent plane and normal to the surfaces:

44. $z = x^2 + 2y^2$ at the point $(1, 1, 3)$.

45. $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$ at the point (x_0, y_0, z_0).

46. $z = \sin x \cos y$ at the point $(\pi/4, \pi/4, 1/2)$.

47. $z = x^2 + y^2 + 2xy$ at the point $(1, 1, 4)$.

48. $x^2 + y^2 + xyz - 3 = 0$ at the point $(1, 1, 1)$.

49. Form the equations of the tangent planes to the surface $x^2 + 2y^2 + 3z^2 = 21$ parallel to the plane $x + 4y + 6z = 0$.

Find three or four first terms in the Taylor expansions of the functions:

50. $f(x, y) = e^x \cos y$ in the neighbourhood of the point $(0, 0)$.

51. $f(x, y) = e^x \ln(1 + y)$ in the neighbourhood of the point $(0, 0)$.

52. $f(x, y) = x^y$ in the neighbourhood of the point $(1, 1)$.

53. $f(x, y) = \tan^{-1}\frac{x - y}{1 + xy}$ in the neighbourhood of the point $(0, 0)$.

54. $f(x, y) = e^{x+y}$ in the neighbourhood of the point $(1, -1)$.

Using the definition of the extremum of a function examine for extremum the following functions:

55. $z = 1 - (x - 2)^4 - (y - 3)^4$ at the point $(2, 3)$.

56. $z = (x - 2)^4 + (y - 3)^4$ at the point $(2, 3)$.

57. $z = x^4 - y^4$ at the point $(0, 0)$.

58. $z = \sin^4 x - (y - 1)^4$ at the point (0, 1).

Using the sufficient conditions for the extremum of a function of two variables, examine for extremum the functions:

59. $z = 2y - x^2 - y^2$.

60. $z = x^2 - 2x + y^2$.

61. $z = 2xy - 4x - 2y$.

62. $z = x^3 + 8y^3 - 6xy + 1$.

63. $z = e^{x/2}(x + y^2)$.

64. Find the absolute maximum and minimum of the function $z = x^2 - y^2$ on the closed circle $x^2 + y^2 \leqslant 1$.

65. Find the absolute maximum and minimum of the function $z = x^2y(4 - x - y)$ in the triangle bounded by the straight lines $x = 0$, $y = 0$, $x + y = 6$.

66. Find the dimensions of a rectangular open pool having the smallest surface area, if its volume is V.

67. Find the dimensions of a rectangular parallelepiped that for a given total surface S has a maximum volume.

Answers

1. $\begin{cases} 0 \leqslant x \leqslant 2, \\ y \geqslant 0, \end{cases}$ and $\begin{cases} -2 \leqslant x \leqslant 0, \\ y \leqslant 0. \end{cases}$ **2.** The square formed by the segments of the straight lines $x = \pm 1$ and $y = \pm 1$, including its sides. **3.** The family of concentric circles $2\pi k \leqslant x^2 + y^2 \leqslant (2k + 1)\pi$, $k = 0, 1, 2, \ldots$. **4.** The entire plane save for the points on the straight lines $y = x$ and $y = 0$. **5.** The part of the plane above the parabola $y = -x^2$. **6.** The points on the circle $x^2 + y^2 = R^2$. **7.** The entire plane save for the straight lines $x + y = n$, $n = 0, \pm 1, \pm 2, \ldots$.

8. (a) $\begin{cases} \sin x \geqslant 0, \\ \sin y \geqslant 0, \end{cases}$ which yields

$$\begin{cases} 2k\pi \leqslant x \leqslant (2k + 1)\pi, & k = 0, \pm 1, \pm 2, \ldots, \\ 2m\pi \leqslant y \leqslant (2m + 1)\pi, & m = 0, \pm 1, \pm 2, \ldots . \end{cases}$$

or $\begin{cases} \sin x \leqslant 0, \\ \sin y \leqslant 0, \end{cases}$ which yields

$$\begin{cases} (2k - 1)\pi \leqslant x \leqslant 2k\pi, & k = 0, \pm 1, \pm 2, \ldots, \\ (2m - 1)\pi \leqslant y \leqslant 2m\pi, & m = 0, \pm 1, \pm 2, \ldots, \end{cases}$$

The domain is the hatched squares (Fig. 12.25)

(b) $\begin{cases} \sin x - 1 = 0, \\ \sin y - 1 = 0, \end{cases}$ which yields

$$\begin{cases} x_k = \dfrac{\pi}{2} + 2k\pi, & k = 0, \pm 1, \pm 2, \ldots, \\ y_m = \dfrac{\pi}{2} + 2m\pi, & m = 0, \pm 1, \pm 2, \ldots . \end{cases}$$

The function is defined at the points $M_{km} = (x_k, y_m)$. **9.** (a) Straight lines parallel to the line $x + y = 0$; (b) concentric circles with centre at the origin of coordinates. **10.** (a) Parabo-

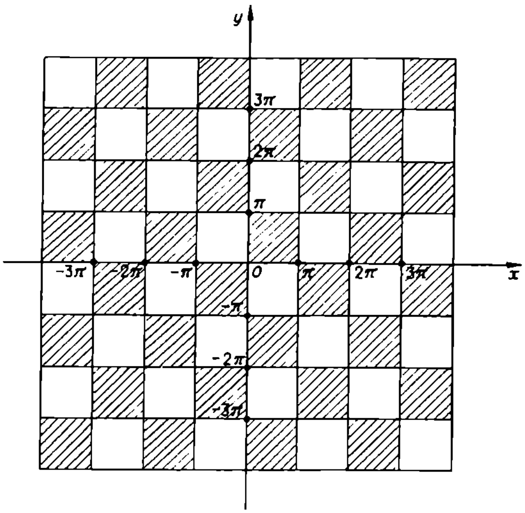


Fig. 12.25

las $y = Cx^2$; (b) parabolas $y = C\sqrt{x}$. **11.** (a) Parabolas $y = C - x^2$ ($C > 0$); (b) Hyperbolas $xy = C$, where $|C| \leqslant 1$. **12.** Planes parallel to the plane $x + y + z = 0$. **13.** For $u > 0$ one-sheet hyperboloids of revolution about the z-axis; for $u < 0$ two-sheet hyperboloids of revolution about the z-axis; both families of surfaces are separated by the cone $x^2 + y^2 - z^2 = 0$. **14.** (a) $-1/4$; (b) 0. **15.** (a) 1; (b) 2. **16.** (a) e^k; (b) 0. **17.** (a) no limit; (b) 0. **18.** We put $y = kx$, then $z = \frac{(1 + k)x}{(1 - k)x} = \frac{1 - k}{1 + k}$, $x \neq 0$. For $k = -1$ we have $\lim\limits_{x \to 0} z = 0$, for $k = 1/2$ $\lim\limits_{x \to 0} z = 3$, and for $k = 3$ $\lim\limits_{x \to 0} z = -2$, so that the given function has no limit at the point (0, 0). **19.** (a) The point (0, 0); (b) the point (0, 0). **20.** (a) The discontinuities form the circle $x^2 + y^2 = 1$; (b) the discontinuities form the line $y = x$. **21.** (a) The points of discontinuities lie on the x- and y-axes; (b) $\varnothing$ (empty set). **22.** All points (m, n), where m and n are integers. **23.** $\frac{\partial z}{\partial x} = 3x^2 - 2y$; $\frac{\partial z}{\partial y} = 2y - 2x$; $dz = (3x^2 - 2y)\,dx + 2(y - x)\,dy$. **24.** $\frac{\partial z}{\partial y} = \frac{y}{x^2 + y^2}$; $\frac{\partial z}{\partial y} = -\frac{x}{x^2 + y^2}$; $dz = \frac{y\,dx - x\,dy}{x^2 + y^2}$. **25.** $\frac{\partial z}{\partial x} = -\frac{1}{y}e^{-x/y}$; $\frac{\partial z}{\partial y} = \frac{x}{y^2}e^{-x/y}$; $dz = \frac{e^{-x/y}}{y^2}(-y\,dx + x\,dy)$. **26.** $\frac{\partial z}{\partial x} = \frac{1}{x + \ln y}$; $\frac{\partial z}{\partial y} = \frac{1}{y(x + \ln y)}$; $dz = \frac{y\,dx + dy}{y(x + \ln y)}$. **27.** $\frac{\partial u}{\partial x} = y + z$; $\frac{\partial u}{\partial y} = x + z$; $\frac{\partial u}{\partial z} = x + y$; $du = (y + z)\,dx + (x + z)\,dy + (x + y)dz$. **28.** $\frac{\partial u}{\partial x} = \frac{x}{\sqrt{x^2 + y^2 + z^2}}$; $\frac{\partial u}{\partial y} = \frac{y}{\sqrt{x^2 + y^2 + z^2}}$; $\frac{\partial u}{\partial z} =$

$\dfrac{z}{\sqrt{x^2+y^2+z^2}}$; $du = \dfrac{x\,dx + y\,dy + z\,dz}{\sqrt{x^2+y^2+z^2}}$. **29.** $\dfrac{\partial z}{\partial x} = \sinh(x^2y+\sinh y)\,2xy$; $\dfrac{\partial z}{\partial y} =$ $\sinh(x^2y+\sinh y)(x^2+\cosh y)$; $dz = \sinh(x^2y+\sinh y)[2xy\,dx + (x^2+\cosh y)\,dy]$. **30.** $du = x^{yz-1}\,[yz\,dx + xz\ln x\,dy + xy\ln x\,dz]$. **31.** (a) $4t^3 + 3t^2 + 2t$; (b) $-2\cosh t$. **32.** (a) $\dfrac{\partial z}{\partial x} = e^y$; $\dfrac{dz}{dx} = e^y\left(1 + \dfrac{x}{1+x^2}\right)$; (b) $\dfrac{\partial z}{\partial x} = \dfrac{2x}{x^2-y^2}$; $\dfrac{dz}{dx} = \dfrac{2(x-ye^x)}{x^2-y^2}$. **33.** (a) $\dfrac{\partial z}{\partial u} = 0$; $\dfrac{\partial z}{\partial v} = 1$; (b) $\dfrac{\partial z}{\partial u} = 4u$; $\dfrac{\partial z}{\partial v} = 4v$. **34.** (a) $\dfrac{\partial z}{\partial x} = f'(u)\left[\dfrac{y}{\sqrt{1-x^2y^2}} + \dfrac{1}{y}\right]$; $\dfrac{\partial z}{\partial y} = f'(u)\left[\dfrac{x}{\sqrt{1-x^2y^2}} - \dfrac{x}{y^2}\right]$; (b) $\dfrac{\partial z}{\partial x} = f'(u)\left[\cos\dfrac{x}{y}\cdot\dfrac{1}{y} + e^{\tan xy}\dfrac{1}{\cos^2 xy}\,y\right]$, $\dfrac{\partial z}{\partial y} = f'(u)\left[-\cos\dfrac{x}{y}\cdot\dfrac{x}{y^2} + e^{\tan xy}\dfrac{1}{\cos^2 xy}\,x\right]$. **35.** (a) $\dfrac{\partial z}{\partial x} = \dfrac{\partial f}{\partial u}\,2xy + \dfrac{\partial f}{\partial v}\,\dfrac{1}{\sqrt{y^2-x^2}}$, $\dfrac{\partial z}{\partial y} = \dfrac{\partial f}{\partial u}\,x^2 - \dfrac{\partial f}{\partial v}\,\dfrac{1}{\sqrt{y^2-x^2}}\,\dfrac{x}{y}$; (b) $\dfrac{\partial z}{\partial x} = \dfrac{\partial f}{\partial u}\,e^{x^2+\cos y}\,2x - \dfrac{\partial f}{\partial v}\,\dfrac{y}{x^2+y^2}$; $\dfrac{\partial z}{\partial y} =$ $-\dfrac{\partial f}{\partial u}\,e^{x^2+\cos y}\cdot\sin y + \dfrac{\partial f}{\partial v}\cdot\dfrac{x}{x^2+y^2}$ **36.** $y' = -\dfrac{x}{y}$. **37.** $y' = \dfrac{y}{x}$. **38.** $y' =$ $\dfrac{(2xy\sqrt{y^2-x^2}+1)\,y^2}{(1-x^2y^2)\sqrt{y^2-x^2}+xy}$. **39.** $y' = \dfrac{yx^{y-1} - y^x\ln y}{xy^{x-1} - x^y\ln x}$. **40.** At the point $M_1(-3;\ 1)$, $y' = 3/4$; at the point $M_2(3;\ 9)$, $y' = -3/4$. **41.** $M_1(-1;\ 3)$, $M_2(-1;\ -1)$. **42.** $\dfrac{\partial z}{\partial x} =$ $\dfrac{z\sin x - \cos y}{\cos x - y\sin z}$; $\dfrac{\partial z}{\partial y} = \dfrac{x\sin y - \cos z}{\cos x - y\sin z}$. **43.** $\dfrac{\partial z}{\partial x} = -\dfrac{c^2x}{a^2z}$; $\dfrac{\partial z}{\partial y} = -\dfrac{c^2y}{b^2z}$. **44.** $2x + 4y - z = 3$; $\dfrac{x-1}{2} = \dfrac{y-1}{4} = \dfrac{z-3}{-1}$. **45.** $\dfrac{xx_0}{a^2} + \dfrac{yy_0}{b^2} + \dfrac{zz_0}{c^2} = 1$; $\dfrac{x-x_0}{\dfrac{x_0}{a^2}} = \dfrac{y-y_0}{\dfrac{y_0}{b^2}} = \dfrac{z-z_0}{\dfrac{z_0}{c^2}}$. **46.** $x-y-2z+1=0$; $\dfrac{x-\pi/4}{1} = \dfrac{y-\pi/4}{-1} = \dfrac{z-1/2}{-2}$. **47.** $4x+4y-z-4=0$; $\dfrac{x-1}{4} = \dfrac{y-1}{4} = \dfrac{z-4}{-1}$. **48.** $3x + 3y + z - 7 = 0$; $\dfrac{x-1}{3} = \dfrac{y-1}{3} = \dfrac{z-1}{1}$.

49. $x + 4y + 6z + 21 = 0$; $x + 4y + 6z - 21 = 0$. **50.** $1 + x + \dfrac{1}{2}(x^2-y^2) + \dfrac{1}{6}(x^3 - 3xy^2)$.

51. $y + \dfrac{1}{2!}(2xy - y^2) + \dfrac{1}{3!}(3x^2y - 3xy^2 + 2y^3)$. **52.** $1 + (x-1) + (x-1)(y-1) + \dfrac{1}{2}(x-1)^2(y-1)$. **53.** *Hint:* Use the formula $\tan^{-1}\dfrac{x-y}{1+xy} = \tan^{-1}x - \tan^{-1}y$. We obtain: $x - y - \dfrac{1}{3}(x^3 - y^3) + \dfrac{1}{5}(x^5 - y^5)$. **54.** $1 + [(x-1)+(y+1)] + \dfrac{[(x-1)+(y+1)]^2}{2!} +$

$\frac{[(x-1)+(y+1)]^3}{3!} = 1 + \frac{x+y}{1!} + \frac{(x+y)^2}{2!} + \frac{(x+y)^3}{3!}$. **55.** $z_{\max} = 1$. **56.** $z_{\min} = 0$. **57.** No extremum. **58.** No extremum. **59.** $z_{\max} = 1$ at the point (0, 1). **60.** $z_{\min} = -1$ at the point (1, 0). **61.** No extremum. **62.** $z_{\min} = 0$ at point (1, 1/2). **63.** $z_{\min} = -\frac{2}{e}$ at point (−2, 0). **64.** The absolute maximum $z = 1$ at the points (1, 0) and (−1, 0). The absolute minimum $z = -1$ at the points (0, 1) and (0, −1). **65.** The absolute maximum $z = 4$ at the point (2, 1). The absolute minimum $z = -64$ at the point (4, 2). **66.** $x = \sqrt[3]{2V}$, $y = \sqrt[3]{2V}$, $z = \frac{1}{2}\sqrt[3]{2V}$. **67.** Cube with a side $a = \sqrt{\frac{S}{6}}$.

Appendix I

Elementary Functions

1. Power functions. The *power function* $y = x^\alpha$, where α is any real number, is defined for all $x > 0$; this function monotonically increases if $\alpha > 0$ and monotonically decreases if $\alpha < 0$ as shown in Figs. I.1 and I.2, respectively.

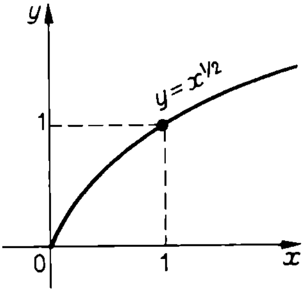


Fig. I.1

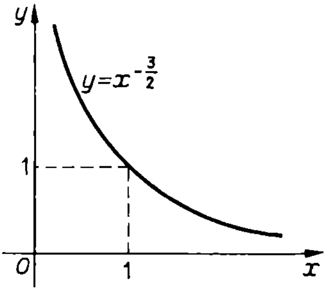


Fig. I.2

If α is a positive integer the function $y = x^\alpha$ is defined at every point of the number line $-\infty < x < +\infty$. The graphs of this function for $\alpha = 3$ and $\alpha = 4$ are shown in Figs. I.3 and I.4, respectively.

If α is a negative integer the function $y = x^\alpha$ is defined for all values of x except $x = 0$ (Figs. I.5 and I.6).

If $\alpha = \frac{p}{q} > 0$, where q is odd, is a rational number the function $y = x^\alpha$ is defined everywhere on the number line; if q is even this function is defined for all $x \geqslant 0$.

2. Exponential functions. The *exponential function* $y = a^x$, where $a > 0$ and $a \neq 1$, is defined at every point of the number line R. The number a is called the base of the power function. This function monotonically increases if $a > 1$ and monotonically decreases if $0 < a < 1$ (Fig. I.7).

3. Logarithmic functions. The *logarithmic function* $y = \log_a x$ to the base a, where $a > 0$ and $a \neq 1$, is defined of the interval $(0, +\infty)$. This

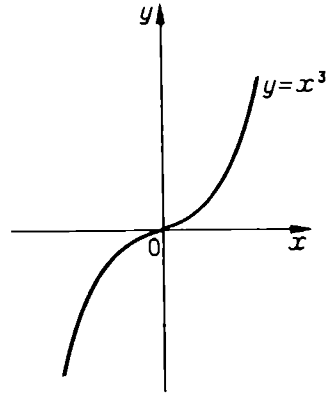


Fig. I.3

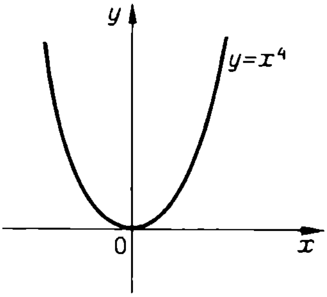


Fig. I.4

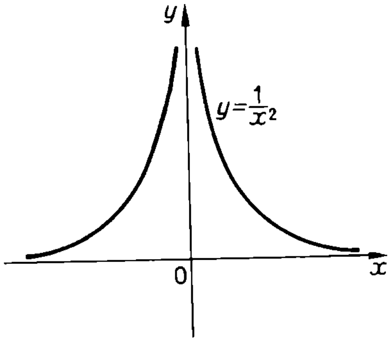


Fig. I.5

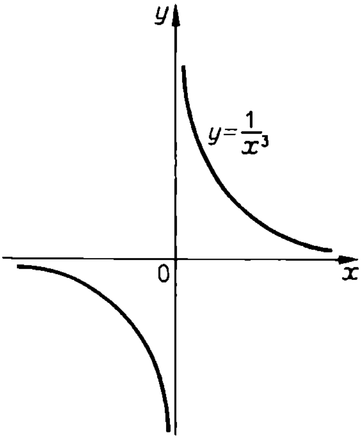


Fig. I.6

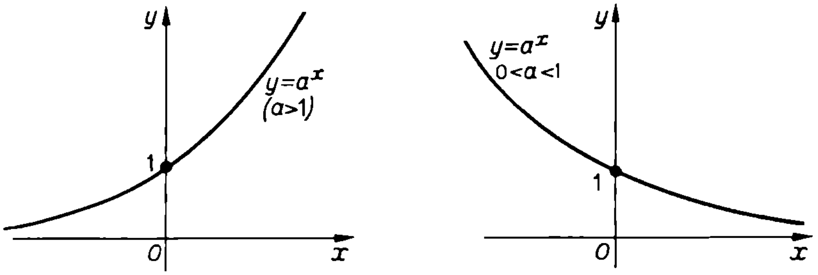


Fig. I.7

function monotonically increases if $a > 1$ and monotonically decreases if $0 < a < 1$ (Fig. I.8).

The logarithmic function $y = \log_a x$ is the inverse of the exponential function $y = a^x$ and vice versa.

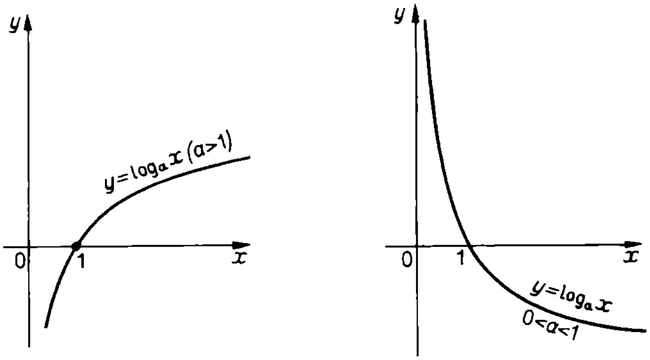


Fig. I.8

The logarithmic function to the base $a = \mathrm{e}$ is called the *natural logarithm* and is usually denoted by $\ln x$; the logarithmic function to the base $a = 10$ is called the *common logarithm* and sometimes denoted by $\log x$. In short $\log_e x = \ln x$ and $\log_{10} x = \log x$.

4. Trigonometric functions. The *sine function* $y = \sin x$ is a periodic function with period $T = 2\pi$ defined for all x. The graph of the sine function is shown in Fig. I.9.

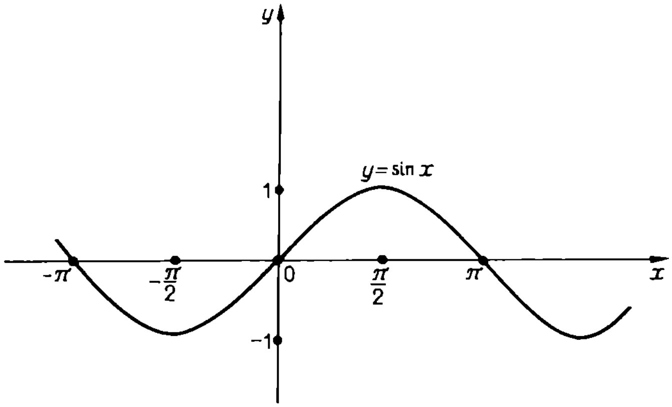


Fig. I.9

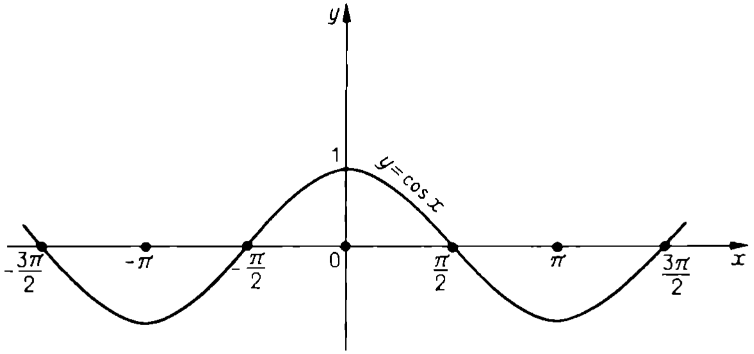


Fig. I.10

The *cosine funcion* $y = \cos x$ is a periodic function with period $T = 2\pi$ defined for all x (Fig. I.10). The graph of this function is obtained from that of the sine function by translating the latter by $-\frac{\pi}{2}$ along the x-axis.

The *tangent function* $y = \tan x$ is a periodic function with period $T = \pi$ defined everywhere except at the points $x = (2k + 1)\frac{\pi}{2}$ $(k = 0, \pm 1, \pm 2, \ldots)$ (Fig. I.11).

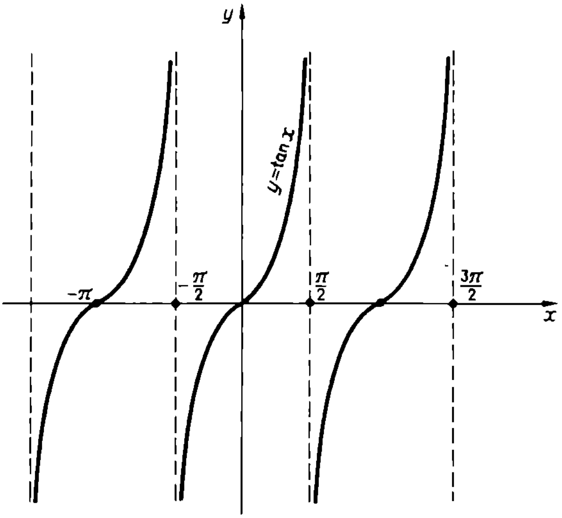


Fig. I.11

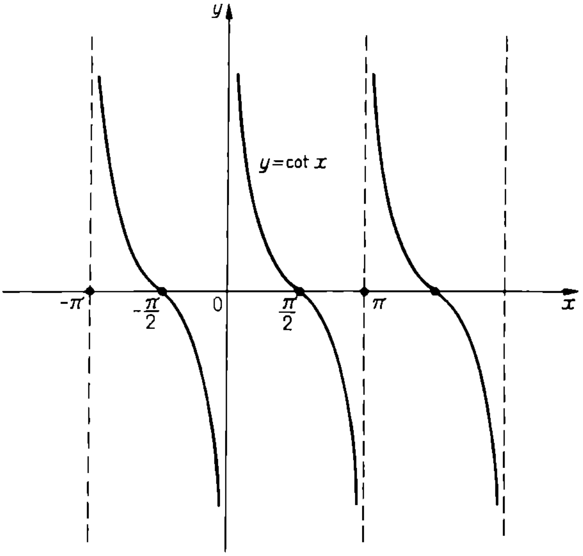


Fig. I.12

The *cotangent function* $y = \cot x$ is a periodic function with period $T = \pi$ defined everywhere except at the points $x = k\pi$ ($k = 0, \pm 1, \pm 2, \ldots$) (Fig. I.12).

The *secant* and the *cosecant functions* are given by the formulas $\sec x = \dfrac{1}{\cos x}$ and $\operatorname{cosec} x = \dfrac{1}{\sin x}$, respectively; they are defined everywhere except at the points where the denominators vanish.

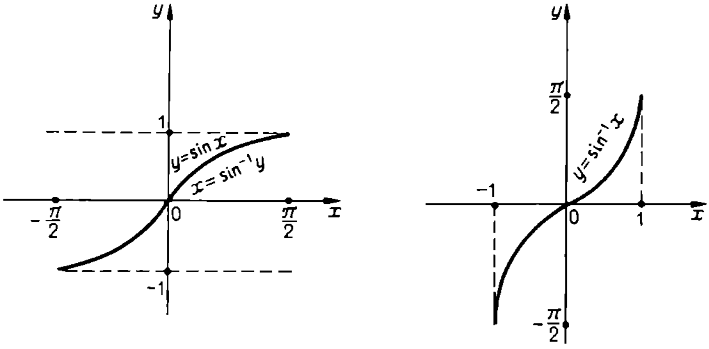


Fig. I.13

5. Inverse trigonometric functions. (1) $y = \sin^{-1} x$.

Let us consider the sine function $y = \sin x$ on the closed interval $[-\pi/2, \pi/2]$, where this function monotonically increases. Hence, the sine function has the inverse function $x = \sin^{-1} y$ defined on the closed interval $[-1, 1]$ so that the range of $x = \sin^{-1} y$ is the closed interval $[-\pi/2, \pi/2]$. The graph of $y = \sin^{-1} x$ is shown in Fig. I.13.

(2) $y = \cos^{-1} x$.

Let us consider the cosine function $y = \cos x$ on the closed interval $[0, \pi]$, where this function monotonically decreases. On the closed interval $[-1, 1]$ there exists the inverse function $x = \cos^{-1} y$ whose values fill in the closed interval $[0, \pi]$. The graph of $y = \cos^{-1} x$ is shown in Fig. I.14.

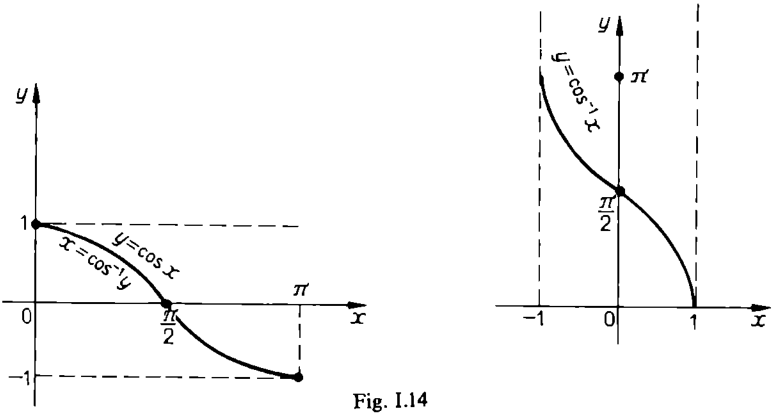


Fig. I.14

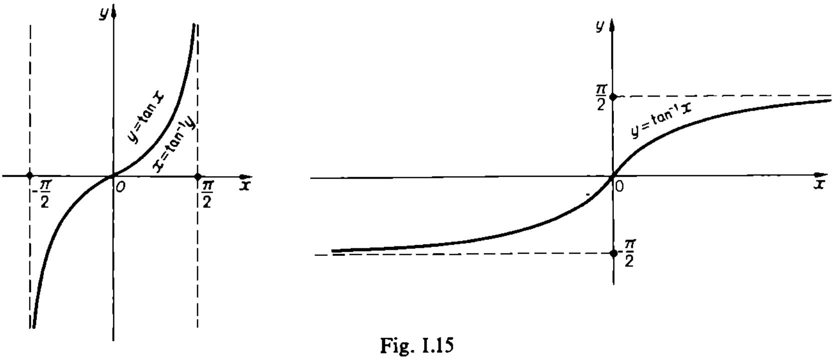


Fig. I.15

(3) $y = \tan^{-1} x$.

Let us consider the tangent function $y = \tan x$ on the open interval $(-\pi/2, \pi/2)$, where this function monotonically increases. Then the range of $y = \tan x$ is the interval $(-\infty, +\infty)$ so that there exists the inverse function $x = \tan^{-1} y$ defined at every point of the number line; the values of $x = \tan^{-1} y$ fill in the interval $[-\pi/2, \pi/2]$. The graph of $y = \tan^{-1} x$ is shown in Fig. I.15.

(4) $y = \cot^{-1} x$.

As before we consider the cotangent function $y = \cot x$ on the interval $(0, \pi)$, where $y = \cot x$ monotonically decreases. The range of $y = \cot x$ is the interval $(-\infty, +\infty)$ so that there exists the inverse function $x = \cot^{-1} y$ defined at every point of the number line; the values of $x = \cot^{-1} y$ fill in the interval $(0, \pi)$. The graph of $y = \cot^{-1} x$ is shown in Fig. I.16.

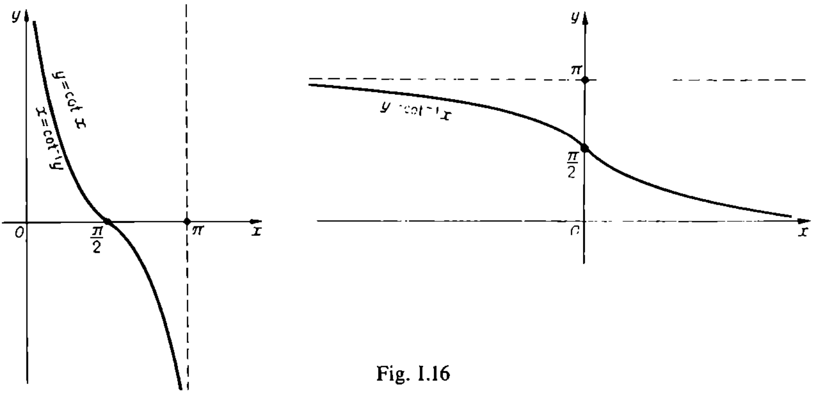


Fig. I.16

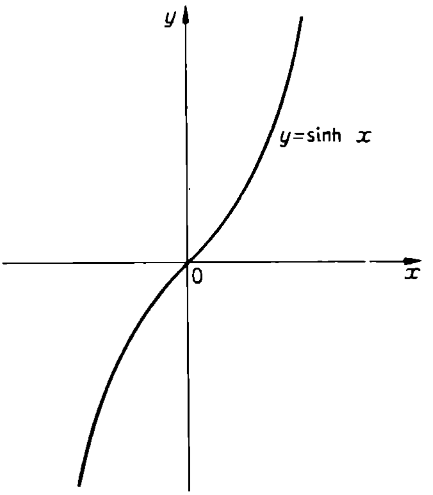


Fig. I.17

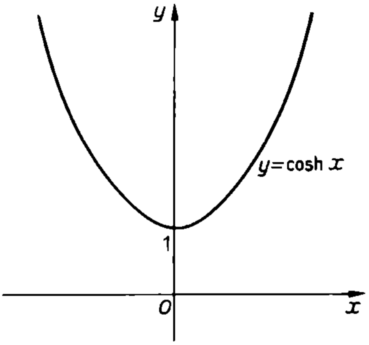


Fig. I.18

6. Hyperbolic functions. The *hyperbolic sine* is given by

$$\sinh x = \frac{e^x - e^{-x}}{2}. \tag{I.1}$$

The domain is the interval $(-\infty, \infty)$.

The range is the interval $(-\infty, \infty)$.

The hyperbolic sine is an odd function since $\sinh(-x) = -\sinh x$. The graph is shown in Fig. I.17.

The *hyperbolic cosine* is given by

$$\cosh x = \frac{e^x + e^{-x}}{2}. \tag{I.2}$$

The domain is the interval $(-\infty, \infty)$.

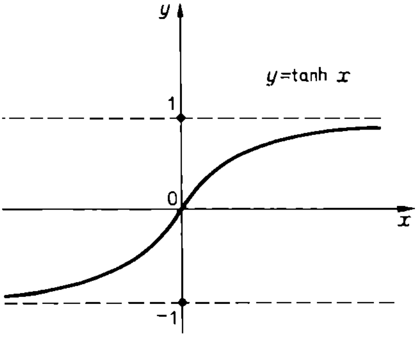


Fig. I.19

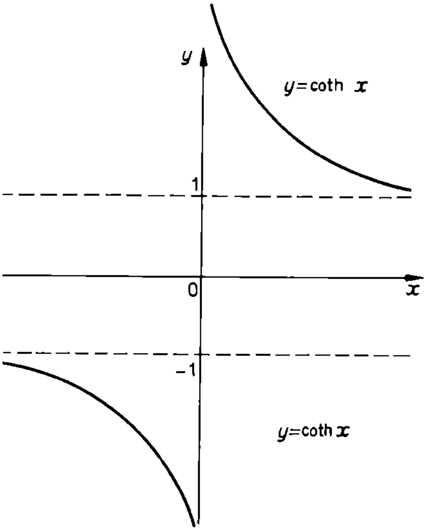


Fig. I.20

The range is the interval $[1, \infty)$.

The hyperbolic cosine is an even function assuming its minimal value at the point $x = 0$. The graph is shown in Fig. I.18.

The *hyperbolic tangent* is given by

$$\tan x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{I.3}$$

The domain is the interval $(-\infty, \infty)$.

The range is the interval $(-1, 1)$ so that $|\tanh x| < 1$.

The hyperbolic tangent is an odd function. The graph is shown in Fig. I.19.

The *hyperbolic cotangent* is given by

$$\coth x = \frac{\cosh x}{\sinh x} = \frac{e^x + e^{-x}}{e^x - e^{-x}}. \tag{I.4}$$

The domain is the union $(-\infty, 0) \cup (0, \infty)$ of two intervals.

The range is the union $(-\infty, -1) \cup (1, \infty)$ of two intervals so that $|\coth x| > 1$.

The hyperbolic cotangent is an odd function. The graph is shown in Fig. I.20.

Relations between hyperbolic functions:

$$\cosh^2 x - \sinh^2 x = 1,$$

$$\cosh x = \sqrt{\sinh^2 x + 1}, \quad \sinh x = \sqrt{\cosh^2 x - 1}.$$

$$\sinh(x + y) = \sinh x \cosh y + \cosh x \sinh y,$$

$$\sinh 2x = 2 \sinh x \cosh x,$$

$$\cosh(x + y) = \cosh x \cosh y + \sinh x \sinh y,$$

$$\cosh 2x = \cosh^2 x + \sinh^2 x.$$

$$\tanh(x + y) = \frac{\tanh x + \tanh y}{1 + \tanh x \tanh y},$$

$$\tanh 2x = \frac{2 \tanh x}{1 + \tanh^2 x}.$$

These relations follow from the formulas (I.1)-(I.4).

Since $\sinh x$ and $\tanh x$ are odd functions the above relations yield

$$\sinh(x - y) = \sinh x \cosh y - \cosh x \sinh y$$

$$\cosh(x - y) = \cosh x \cosh y - \sinh x \sinh y$$

$$\tanh(x - y) = \frac{\tanh x - \tanh y}{1 - \tanh y \tanh y}.$$

In conclusion we shall prove the relation

$$(\sinh x + \cosh x)^n = \sinh nx + \cosh nx \quad (n \in N).$$

Indeed, using (I.1) and (I.2), we get

$$(\sinh x + \cosh x)^n = \left(\frac{e^x - e^{-x}}{2} + \frac{e^x + e^{-x}}{2}\right)^n = (e^x)^n = e^{nx}$$

$$= \frac{e^{nx} - e^{-nx}}{2} + \frac{e^{nx} + e^{-nx}}{2} = \sinh nx + \cosh nx.$$

Index

Abscissa, 14
Absolute value, 231
Adjoint operator, 210
  properties of, 210-211
Angle
  dihedral, 50
  polar, 18
Antiderivative, 409
Applicate, 14
Approximation
  accuracy of, 232
  parabolic, 498
  precision of, 232
  Simpson's, 502
  trapezoidal, 498
Area
  of a curvilinear figure, 457
  of a plane figure, 476
Arc, 488
Asymptotes
  horizontal, 372
  oblique, 369
  vertical, 367
Axis, 11
  coordinate, 11
  of abscissas, 12
  of applicates, 13
  of ordinates, 12
  of parabola, 78
  polar, 18
Axes of coordinates
  reflection of, 65
  rotation of, 64
  translation of, 63

Basis
  of linear space, 175
  orthonormal, 31
Bernoulli's inequality, 238
Bilinear form, 214
Boundary point, 531
Bounded set, 235

Cartesian coordinates, 11
  in a plane, 12
  in three-dimensional space, 13
  origin of, 13
Cartesian coordinate system, 14
  compatible with a polar system, 18
Cauchy-Schwarz inequality, 183
Circle, 67
  equation of, 15
Cofactor, 126
Column-vector, 103
Complex number(s), 294
  exponential form of, 299
  extracting a root of, 299
  operations on, 295
  representation of, 294
  trigonometric form of, 296
Components of a vector, 30
Composite function, 278
  derivatives of, 322, 545
  differentials of, 323, 549
Conditions
  necessary, 237
  sufficient, 237
Cone of the second order, 102
Conic surface
  directing line of, 93
  generating line of, 93
  vertex of, 93
Convexity of a curve, 362
Coordinate axis, 11
  reference point on, 11
  unit distance on, 11
Coordinate plane, 14
Coordinate system
  Cartesian, 14
  polar, 18
  rectangular, 12
Coordinates
  of a vector, 30
  origin of, 12
  polar, 18
Cramer's rule, 154
Curve(s) of the second order, 66
  classification of, 83-89
  equation of, 67
  standard equations of, 85-89
Cylinder
  elliptic, 93, 101
  hyperbolic, 101
  parabolic, 101
Cylindrical surface
  directing line of, 91
  generating line of, 91

Derivative(s)
  geometric interpretation of, 306
  infinite, 309
  left-hand, 309
  of a constant function, 316
  of elementary functions, 319
  of higher orders, 332, 559
  of a ptocuct of functions, 317
  of a quotient of functions, 318
  of a sum of functions, 316
  partial, 538
  right-hand, 309
Determinant(s)
  computing, 132-133
  multiplication of, 130
  of a matrix of order *n*, 124
  of a square matrix, 122
  of a triangular matrix, 125
  principal diagonal of, 19
  properties of, 129-132
  second-order, 19
  secondary diagonal of, 19
  third-order, 20
  transposition of, 130
Differentiable functions, 312
  continuity of, 313
Differential(s), 314
  invariance of form of, 323
  of a composite function, 549
  of higher orders, 334, 559
  partial, 545
  total, 544
Directed line, 11
Directed segment
  magnitude of, 33
Direction cosines, 37
Dirichlet's function, 460

Discontinuity
  removable, 281
  unremovable, 282

Elementary functions
  continuity of, 275
  derivatives of, 319
Ellipse, 67
  directrices of, 70
  eccentricity of, 70
  equation of, 16, 86
  foci of, 68
  imaginary, 86
  optical propery of, 81-83
  properties of, 68-71
  standard Cartesian equation of, 67
  vertices of, 68
Ellipsoid, 95
  of revolution, 95
Elliptic cylinder, 93
Equation(s)
  approximate solution of, 381
  constraint, 573
  of a circle, 15
  of a curve of the second
  of an ellipse, 16, 86
  order, 67
Equation of a plane
  general, 48
  normal coordinate, 49
  normal vector, 48
Error
  absolute, 232
  relative, 232
Extremum of a function, 355
Euclidean space(s)
  angle between vectors in, 184
  examples of, 183
  length of a vector in, 184
  orthogonalization in, 185-189
  orthonormal basis of, 187
  orthonormal vectors in, 185
  real, 182

Fixed vector(s), 24
  as a directed segment, 24
  equivalent, 24
  zero, 24
Free vector(s), 24
  collinear, 28
  length of, 25
Function(s)
  conditional extrema of, 573
  composite, 278, 545
  continuous on closed intervals, 283
  continuous on a domain, 537
  continuous at a point, 272, 536
  critical points of, 568
  derivatives of higher orders of, 558
  differentiable, 541
  differentials of higher orders of, 559
  discontinuities of, 280
  domain of, 247
  elementary, 275
  exponential, 587
  extrema of, 352, 566
  homogeneous of degree q, 93
  hyperbolic, 594
  implicit, 550
  infimum of, 286
  integrable, 460
  inverse trigonometric, 592
  limit of, 533
  linear combination of, 412
  logarithmic, 587
  Maclaurin's formula for, 564
  of one variable, 247
  of several variables, 305
  partial derivatives of, 538
  power, 587
  range of, 247
  rational, 424
  representation of, 249-251
  smooth, 312
  supremum of, 286
  Taylor's formula for, 564
  trigonometric, 589
Fundamental theorems of calculus
  first, 469
  second, 470

Gaussian elimination, 61-63
General equation of a plane, 48

Homogeneous linear system(s)
  fundamental system of solutions of, 160
  properties of, 156-157
L'Hospital's rule, 345
Hyperboloid
  of one sheet, 96
  of two sheets, 97
Hyperboloid of revolution
  one-sheet, 96
  two-sheet, 97
Hyperbola, 71
  asymptotes of, 72
  branches of, 72
  conjugate, 77
  directrices of, 75
  eccentricity of, 75
  equation of, 86
  foci of, 74
  optical property of, 83
  properties of, 71-77
  standard Cartesian equation of, 71
  vertices of, 71

Implicit function, 550
Improper integral(s)
  Cauchy principal value of, 519
  Dirichlet's convergence test for, 518
  of unbounded functions, 520
  of unbounded nonnegative functions, 523
Indefinite integral(s), 410
  involving elementary functions, 412
  properties of, 411
Inequality
  Bernoulli's, 238
  Cauchy-Schwarz, 183
  triangle, 184
Infinitesimal(s), 258
  asymptotically equal, 290
  comparison of, 288
  equivalent, 289
  properties of, 259
Infinity, 262
Integral(s)
  definite, 459
  elliptic, 436
  improper, 507
  indefinite, 410
  involving irrational functions, 435
  involving trigonometric functions, 445
  of nonnegative functions, 511
Integration
  numerical, 498
  by parts, 417, 475
  by substitution, 414, 472
  of partial fractions, 429
  of rational functions, 424
Interval
  closed, 234
  half-closed, 234
  half-open, 234
  infinite, 234
  open, 234
Inverse functions, 324
  differentiation of, 324
Irrational functions, 435

Kronecker-Capelli theorem, 145

Landau symbols, 293
Limit(s) of a function
  Cauchy criterion for, 251
  geometric interpretation of, 251

Limit(s) of a function
  left-hand, 270
  operations on, 266
  right-hand, 270
  sequential criterion for, 253
Limit of a sequence of complex numbers, 301
Limit theorems, 254-256
Line(s), 48
  equation of, 52
  direction numbers of, 55
  direction vector of, 55
  general equation of, 51, 56
  imaginary, 86
  intercept form of equation of, 53
  normal coordinate equation of, 51
  normal vector equation of, 51
  slope intercept form of
  parametric vector equation of, 55
  point direction equation of, 56
Linear mapping(s), 192
  examples of, 192-193
  image of, 193
  kernel of, 196
  nullity of, 196
  operations on, 196-197
  product of, 196
  rank of, 194
  sum of, 196
Linear operator(s), 197
  adjoint operator of, 209
  characteristic polynomial of, 205
  eigenvalues of, 205
  eigenvectors of, 205
  inverse of, 198
  matrix of, 200
  multiplication of, 197
  symmetric (self-adjoint), 211
Linear space(s), 168
  additive inverse of a vector in, 168
  basis of, 175
  dimension of, 178
  examples of, 168-170
  linear subspace in, 170
  linearly dependent vectors in, 174
  linearly independent vectors in, 174
  n-dimensional, 180
  properties of, 170
  vectors in, 168
  zero vector in, 168
Linear span(s), 172
  examples of, 173-174
  properties of, 173
Linear subspace(s), 170
  direct sum of, 172
  examples of, 170-171
  intersection of, 172
  orthocompliment of, 189
  sum of, 171
Logical connectives, 236-237

Maclaurin's formula, 389, 564
  for elementary functions, 390
Mapping(s)
  identity, 193
  linear, 192
Mathematical induction, 238
  method of, 238
  principle of, 238
Matrix(ces), 103
  addition of, 105
  base column of, 140
  base minor of, 140
  base row of, 140
  elementary, 119
  elementary operations on, 113-122
  identity, 104
  inverse of, 134
  minor of, 125
  multiplication of, 105
  nonsingular, 133
  of linear operator, 200
  of schematic form, 116
  principal diagonal of, 104
  product of, 105
  rank of, 140
  square, 122
  transition, 182
  transpose of, 110
  transposition of, 110
  triangular, 124
  unit, 104
  zero, 124
Method of simple iteration, 163-165
Methods of solving systems of linear equations
  direct, 161
  indirect, 161
  iterative, 161
Mixed product of vectors, 43
Multiplication of matrices, 106-109
  laws of, 108

Newton-Leibniz theorem, 471
Number(s)
  irrational, 231
  natural, 230
  rational, 230
  real, 230
Number line, 234

Octant, 14
Operator
  adjoint, 209
  linear, 197
  symmetric, 211
Ordinate, 14
Orthocompliment of linear subspace, 189
  properties of, 189-190
Orthonormal basis, 31

Parabola, 77
  axis of, 78
  directrix of, 78
  focal parameter of, 78
  focus of, 78
  optical property of, 83
  properties of, 77-81
  standard Cartesian equation of, 77
  tangent to, 80
  vertex of, 77
Parabolic approximation, 498
Paraboloid
  elliptic, 98
  hyperbolic, 98
  of revolution, 97
Partial derivative(s), 538
  geometric interpretation of, 540
  mixed, 559
Partial fraction, 425
Plane, 47
  general equation of, 48
  normal coordinate equation of, 49
  normal vector equation of, 47
Plane curve
  equation of, 66
Point
  boundary, 531
  critical (stationary), 568
  deleted δ-neighbourhood of, 235
  δ-neighbourhood of, 235
  of inflection, 362
  of local maximum, 353
  of local minimum, 353
  of strict maximum, 354
  of strict minimum, 354
  regular (nonsingular), 555
  singular, 555
Polar axis, 18
Polar coordinate system, 18
Polar coordinates
  polar angle of, 18
  pole of, 18
Polar radius, 18
Position vector, 31

Quadratic form, 213
  associated matrix of, 213
  bilinear form associated with, 214
  diagonalization of, 219
  law of inertia for, 221
  positive-definite, 219
Quantifier
  existential, 236
  universal, 236

Rational function, 424
  proper, 425
  real, 425
Reflection of axes of coordinates, 65
Relations between Cartesian and polar coordinates, 18
Relations between infinitesimals and infinities, 265
Rotation of axes of coordinates, 64
Row-vector(s), 103
  linear combination of, 112
  linear dependence of, 112
  linearly independent, 112
  nontrivial linear combination of, 112
  trivial linear combination of, 112

Space
  n-dimensional real coordinate, 169
  Euclidean, 183
  linear, 168
  linear complex, 168
  linear real, 168
  unitary, 191
  vector, 168
Standard Cartesian coordinate system, 67
Surface(s)
  classification of, 90-95
  conic, 93
  cylindrical, 91
  equation of, 89
  of revolution, 90
  of the second order, 89-90
Surface(s) of the second order
  equation of, 90
  standard equations of, 95-102
Symmetric operator
  properties of, 212-213
System(s) of linear equations, 143
  augmented matrix of, 144
  coefficients of, 144
  coefficient matrix of, 144
  Cramer's rule for, 154
  method of Gaussian elimination for, 148-150
  quadratic, 154
Scalar product, 34
  basic properties of, 34-36
Scalar square of a vector, 36
Scalar triple product, 43
  geometric interpretation of, 43
Second-order determinant, 19
Sequence of numbers, 239
  bounded, 241
  bounded above, 241
  bounded below, 241
  Cauchy convergence criterion of, 241
  infinitely large, 242
  limit of, 239
  stationary, 240
Set(s), 229
  bounded, 235
  bounded above, 235
  bounded below, 235
  connected, 531
  countable, 230
  disjoint, 230
  empty, 230
  equal, 229
  equivalent, 230
  finite, 230
  greatest lower bound of, 236
  infimum of, 236
  infinite, 230
  least upper bound of, 236
  lower bound of, 236
  one-to-one correspondence between, 230
  open, 531
  operations on, 230
  proper subset of, 229
  supremum of, 235
  unbounded above, 235
  unbounded below, 235
  unconnected, 531
  upper bound of, 235
Significant digit, 233
  accurate, 233
Simpson's approximation, 502
Subset, 229
  proper, 229
Taylor's formula, 385, 564
  for functions, 386
  for polynomials, 385
Theorem
  Cauchy mean value, 343
  mean value, 341
  Newton-Leibniz, 471
  on construction of a linear mapping, 191
  Rolle's, 339
Third-order determinant, 20
  cofactor of, 22
  minor of, 21
Total differential
  geometric interpretation of, 557
Transition matrix, 182
  properties of, 182-183
Translation of axes of coordinates, 63
Trapezoidal approximation, 498
Triangle inequality, 184

Unit vector(s), 29
  as orthonormal basis, 31
Unitary space, 191

Vector(s)
  addition of, 26
  direction of, 557
  direction cosines of, 37
  equivalent, 29
  components of, 30
  coordinates of, 30
  coplanar, 43
  fixed, 24
  free, 24
  linear operations on, 26
  moving position, 48
  normal, 49
  orthogonal, 185
  orthonormal, 185
  position, 31
  scalar product of, 34
  scalar square of, 36
  sliding, 25
  sum of, 26
  unit, 29
Vector function, 396
  continuity of, 398
  differentiation of, 399
  limit of, 398
Vector product, 39
Vector space, 168
Vector triple product, 45

TO THE READER

Mir Publishers would be grateful for your comments on the content, translation and design of this book. We would also be pleased to receive any other suggestions you may wish to make.

Our address is:
Mir Publishers
2 Pervy Rizhsky Pereulok
I-110, GSP, Moscow, 129820
USSR